

C811

F-767

科学专著丛书

抽 样 调 查

—理论、方法与实践

冯士雍 施锡铨 著

上海科学技术出版社

责任编辑 赵序明

科学专著丛书

抽 样 调 查

——理论、方法与实践

冯士雍 施锡铨 著

上海科学技术出版社出版、发行

(上海瑞金二路 450 号)

新华书店上海发行所经销 上海商务印刷厂印刷

开本 787×1092 1/小16 印张 28 插页 4 字数 442,000

1996年8月第1版 1996年2月第1次印刷

印数 1—1,200

ISBN 7-5323-3777-4/O·188

定价: 49.00 元

《科学专著丛书》序

如果说科学研究论文是创造性科学工作的发表性记录,那末科学技术学术专著则是创造性科学工作的总结性记录。前者注重的是优先权,后者注重的是系统化。

在大量科学研究的基础上,对一个专题或一个领域的研究成果,作系统的整理总结,著书立说,乃是科学研究工作不可少的一个组成部分。著书立说,既是丰富人类知识宝库的需要,也是探索未知领域、开拓人类知识新疆界的需要。特别是在科学各门类的那些基本问题上,一部优秀的学术专著常常成为本学科或相关学科取得突破性进展的基石。所以,科学技术学术专著的著述和出版是一项十分重要的工作。

近20年来,中国的科学事业有了迅速的发展,涌现了许多优秀的科学研究成果,为出版学术专著提供了坚实的基础。值此20世纪90年代,在出版学术专著方面,中国的科学界和出版界都在抓紧为本世纪再加些积累,为迎接新世纪多作些开拓。我高兴地看到,作为这种努力的一个部分,《科学》杂志的出版者——上海科学技术出版社推出了这套《科学专著丛书》。

上海科学技术出版社是科学技术界熟悉和信赖的一家出版社,历来注重科学技术学术专著的出版。《科学》杂志的编者组织编辑学术系列丛书,也不是第一次。在本世纪三四十年代,就曾推出过《科学丛书》,其中不乏佳作,对当时的学术研究起了很好的作用。

《科学》在中国是一份历史最长的综合性科学刊物,80年来与科学技术界建立了广泛的密切联系。现在推出的这套《科学专著丛书》正是这种

联系的产物。我相信,加强这种联系,著者与编者、出版者,科技界与出版界共同努力,精心选题,精心编辑,精心出版,一定能使这套专著丛书反映出中国科学技术研究的最新水平,为本世纪多留下几本中国学者的优秀专著,为迈向新世纪多铺下几块引路的基石!

周光召

(《科学》杂志编委会主编)

1994年8月

本书序

1991年夏初,冯士雍教授应中国科学技术大学数学系之聘,为该系统计专门化学生讲授《抽样调查》课程。我当时适因事去合肥,与冯教授朝夕相见,因有幸拜读了他的讲稿,即此书的原胚,披阅之下,深感此稿取材精当,论述严谨而可读性强,尤其难得的是,其中包含了不少实例分析,实是一部极有出版价值的著作。后以此意与冯教授商议,知他也有这个打算,很是高兴。

光阴似箭,很快又过去了三个寒暑,在这期间,曾多次向他打听此书写作的进展情况。承告以此书系我国第一部抽样调查方面的大型著作,故在取材、编排和论述方面务求精当。因此颇费周折,加以科研事务繁重,故对进度有所影响云云。所幸经过几年的努力,这件精雕细刻的工作终于得以呈献于读者之前。作为本书的第一个读者,在感到欣慰之余,也深为作者这种勤勉刻苦、锲而不舍和精益求精的精神所感动。

关于本书的内容与特点,作者在前言中已有充分的介绍。此外想特别提请读者注意的是书中的“案例分析”部分,在其中所汇集的一些大型的抽样调查应用实例,许多是由冯教授主持或参与的,有的并曾在全国性的媒体上报导过,冯教授详细介绍了这些项目的调查目的、方法、指标的选择、调查的组织实施、抽样方案的制定、数据处理方法以及所得结果的解释和应用等的全过程。在某种意义上,这部分可视为冯教授从事抽样调查研究和应用工作十余年的经验的结晶。现在他把这些总结在本书里以与读者共享,实在是弥足珍贵。不列颠百科全书把统计学定义为“收集和分析数据的艺术”。这“艺术”一词值得玩味。而冯教授所提供的这些案例分析,对在抽样调查的领域内这种“艺术”如何展现,提供了感性的范例。其启示的意义,实在是超出这些例子本身之外的。

数理统计学的快速发展,也使抽样调查这门相对说来较为经典的分支学科的面貌有了不少更新。这些在本书中不少地方有所反映。在此还特别要提到由施锡铨教授主持撰写的9、10两章。施教授是我国知名的

中年统计学家，其在 Jackknife 和 Bootstrap 方面的研究在国内堪称独步。由他来承担这类题目的主笔，可说是适人适选。

学术著作出版难，是当前困扰学界同人的一件憾事，而本书这样一部有很大学术和社会意义的著作，在上海科学技术出版社的大力支持下，得以顺利且迅速地问世，其扶持学术的远大眼光功不可没，作为本书读者及学界一员，愿借此机会表示个人的赞赏和钦佩。

陈希孺 1994 年 5 月

前 言

抽样调查历来是应用统计的一个重要分支,在各个领域,特别是社会经济领域中有极其重要的应用,但直到十多年前,在我国应用面还很窄,而在学术界则可说是几乎一片空白。随着我国改革开放的不断深入及社会主义市场经济的初步形成,抽样调查在调查方法中将逐渐占据主导地位,随着它愈来愈广泛的应用,对方法及理论的需求也愈加迫切,但目前国内尚无一本理论、方法与实践这三个方面兼顾的抽样调查的学术著作,本书是为填补这一空缺所作的一个尝试。

由于工作的需要,十余年前,作者开始把注意力转向抽样调查这个领域,除进行理论研究,参加实际项目外,还自1985起先后在中国科技大学、华东师范大学、上海财经大学、上海交通大学、中国科学院研究生院等单位为本科生及硕士研究生开设了《抽样调查》课程,最初所用的讲义在框架上参考了William G. Cochran的*Sampling Techniques*。在本书写作过程中,参照本学科近年发展状况及实际需要,并结合自己一些研究心得,对本书的章节编排及叙述论证作了设计,特别是增加了许多我国的数字实例,其中不少是作者亲自参与的。考虑到新近发展起来的再抽样(resampling)等方法在复杂样本方差估计等方面的理论和实用意义,由在这个领域里富有研究成果的施锡铨教授撰写了第9、10两章,这两章使全书增色不少。

全书共分十一章。第1章介绍抽样调查的意义与作用、若干基本概念及其应用领域,从第2章至第8章详尽地介绍了几种应用中最重要抽样方法:简单随机抽样、分层抽样、不等概率抽样、整群抽样、二阶及多阶抽样、系统抽样,其中第4章介绍两种重要的非线性估计——比估计及回归估计。上述各章的重点是介绍各种抽样方法的适用场合与实施方法、样本量的确定及总体目标量的估计及其方差估计。第9章介绍复杂样本的方差估计方法,第10章讨论了抽样调查的误差来源,特别是非抽样误差及其相关问题的处理方法。最后一章是案例分析,介绍了多项实

际抽样调查项目的背景、目的、设计与分析,并对其进行分析评价。这一章的材料是经过精心选择的,为的是尽可能照顾到不同的应用领域和所用的方法。大部分案例采用作者及其同事们多年来承担的实际项目,另外一些则取自我国开展的其他有影响的全国性抽样调查方案,希望通过这些案例,使读者就一个实际抽样调查项目的目的、设计与分析的全过程有一个概要的了解。

本书主要读者对象是从事抽样调查的理论与方法研究的科研工作者、教师以及实际工作者。为尽可能照顾到多方面读者的需要,本书写作时尽可能做到简繁相间,难易结合。对于那些主要从事实际工作的读者,在初次阅读时可略去若干定理的证明,本书也适合作为研究生的教材、数理统计专业本科生学习抽样调查课程的主要参考书。

本书成稿过程中得到了作者的老师和许多同行的多方鼓励和支持,其中有陈希孺、成平、项可风、张尧庭、汪仁官、孙山泽、菲诗松和梁小筠诸位教授。陈希孺教授在百忙中为本书专门作序,汪仁官教授认真地审阅了全书并提出不少中肯而宝贵的意见。对此作者谨表示深切地感谢。作者还要感谢我的学生邹国华,他十分仔细地阅读了原稿,并对原稿进行了若干校正。最后作者还要特别感谢上海科学技术出版社为编辑出版这本书所作的努力。由于作者水平所限,书中一定存在不足之处,望请有关专家及广大读者惠予批评指正。

冯士雍

1994年4月于中国科学院

系统科学研究所

内 容 提 要

本书所反映的研究工作系国家自然科学基金重点资助项目之一。

全书共分 11 章。第 1 章介绍抽样调查的意义和作用、若干基本概念及其应用范围。从第 2 章至第 8 章详尽地介绍了几种应用中最常用的抽样方法: 简单随机抽样、系统抽样、分层抽样、不等概率抽样、整群抽样、二阶及多阶抽样, 其中第 4 章介绍了两种重要的非线性估计——比估计及回归估计。上述各章的重点在于阐述这些抽样方法的适用场合、实施方法、样本量的确定与总体目标量的估计及方差估计。第 9 章介绍几种复杂样本的方差估计方法, 第 10 章讨论了抽样调查的误差来源, 特别是非抽样误差的处理方法。最后一章案例分析介绍多项实际抽样调查项目的背景、目的、设计与分析, 并对其进行分析评价。

本书主要读者对象是从事抽样调查理论与方法研究的科研人员及实际工作者。也可供数理统计及经济统计专业学生或相关专业研究生作教材或主要参考书。

目 录

《科学专著丛书》序

本书序

前言

第1章 引论	1
§1.1 抽样调查的意义和作用	1
§1.2 若干基本概念	3
1.2.1 总体与样本	3
1.2.2 概率抽样	4
1.2.3 抽样单元与抽样框	5
1.2.4 总体参数的分类	6
1.2.5 误差来源与精度表示	7
1.2.6 样本量、费用与效率	9
§1.3 几种基本的抽样方法	10
1.3.1 简单随机抽样	10
1.3.2 分层抽样	10
1.3.3 整群抽样	11
1.3.4 多阶抽样	11
1.3.5 系统抽样	11
§1.4 抽样调查的步骤	12
§1.5 抽样调查的应用与历史发展	14
1.5.1 主要应用领域	14
1.5.2 国际发展简史	15
1.5.3 我国抽样调查的应用与发展	18
第2章 简单随机抽样	22
§2.1 定义及实施方法	22
2.1.1 从一个有限总体中抽取所有可能的样本	22

2.1.2	两个等价的定义	23
2.1.3	简单随机抽样的实施方法	24
§ 2.2	估计量及其性质	25
2.2.1	简单估计及其无偏性	25
2.2.2	估计量的方差与协方差	28
2.2.3	方差与协方差的估计	31
2.2.4	简单估计的优良性及可以进一步改进的途径	33
■ 2.3	总体比例的估计与对子总体的估计	36
2.3.1	总体比例(百分率)的估计	36
2.3.2	子总体的估计	38
§ 2.4	样本量的确定	41
2.4.1	确定 n 的一般原则	41
2.4.2	总体参数为 Y 或 \bar{Y} 的一般情形	42
2.4.3	估计总体比例 P 的情形	43
2.4.4	总体方差的预先估计	45
§ 2.5	放回简单随机抽样	46
2.5.1	抽样方法及基本特征	46
2.5.2	总体平均数 \bar{Y} 估计量 \bar{y} 的性质	47
2.5.3	设计效应(deff)	49
2.5.4	\bar{Y} 的另一种估计量	49
§ 2.6	利用随机数骰子和随机数表进行随机抽样的方法	50
2.6.1	随机数骰子及其使用方法	50
2.6.2	随机数表的便使用方法	52
第3章	分层抽样	54
§ 3.1	一般描述	54
3.1.1	定义与记号	54
3.1.2	分层抽样适用的场合和优点	55
§ 3.2	估计量及其性质	55
3.2.1	估计量的构造	55
3.2.2	基本性质	56
3.2.3	比例分配及自加权样本	57
3.2.4	一个简单的实验例子	59

§ 3.3 最优分配	60
3.3.1 最优分配的定义	60
3.3.2 主要结果	61
3.3.3 Neyman(最优)分配	62
3.3.4 某些层需要超过 100% 抽样时的修正	64
§ 3.4 分层随机抽样在精度上的得益	64
3.4.1 与简单随机抽样的比较	64
3.4.2 何时分层及最优分配的精度得益最大	65
3.4.3 分层随机抽样精度反比简单随机抽样差的情形	66
3.4.4 从样本估计分层随机抽样精度的得益	68
3.4.5 数值例子——关于职工月平均奖金额的调查	69
3.4.6 偏离最优分配时对方差的影响	72
3.4.7 多指标情形样本量的分配	73
§ 3.5 样本总量 n 的确定	75
3.5.1 估计的总体参数为 \bar{Y} 的情形	75
3.5.2 估计的总体参数为 Y 的情形	78
§ 3.6 对总体比例(百分率)的分层随机抽样	78
3.6.1 估计量及其方差	79
3.6.2 最优分配	79
3.6.3 分层和最优分配精度上的得益	79
3.6.4 样本量的估计	80
§ 3.7 分层技术的充分利用	81
3.7.1 层的构造	81
3.7.2 多重分层	85
3.7.3 每层只抽一个单元时的方差估计	87
3.7.4 事后分层	88
3.7.5 定额抽样	91
§ 3.8 用于分层的二相抽样	91
3.8.1 层权误差对分层估计的影响	91
3.8.2 二相抽样及估计量均值与方差的一般公式	93
3.8.3 用于分层的二相抽样的估计	96

第 4 章 比估计与回归估计	100
----------------------	-----

§ 4.1 比估计及其基本性质	100
4.1.1 定义	100
4.1.2 基本性质	101
4.1.3 方差的估计	103
4.1.4 置信限	104
4.1.5 比估计与简单估计量的比较	105
4.1.6 数值例子 小麦估产调查	105
4.1.7 乘积估计	107
§ 4.2 比估计的偏倚及其均方误差和方差估计的阶	108
4.2.1 关于有限总体样本中心矩阶的基本引理	108
4.2.2 比估计的偏倚与均方误差及其阶的估计	111
4.2.3 均方误差或方差估计的偏倚	115
§ 4.3 分层随机抽样中的比估计	117
4.3.1 分别比估计	118
4.3.2 联合比估计	118
4.3.3 分别比估计与联合比估计的比较	120
4.3.4 分层比估计时的最优分配	121
4.3.5 数值例子——耕地面积核实调查	121
§ 4.4 消除或减少比估计偏倚的方法	124
4.4.1 无偏的比类型估计量	125
4.4.2 减少比估计偏倚的方法	127
§ 4.5 回归估计量 (β 设定时的情形)	129
4.5.1 回归估计量的一般形式	129
4.5.2 β 设定情形的一般结果	128
4.5.3 差估计量	131
§ 4.6 回归估计量 (β 取样本回归系数的情形)	131
4.6.1 表达式及若干引理	131
4.6.2 基本性质	135
4.6.3 回归估计量与简单估计量及比估计量的比较	138
§ 4.7 分层随机抽样中的回归估计	139
4.7.1 分别回归估计	139
4.7.2 联合回归估计	140
4.7.3 数值例子——专业技术人员总数的调查	142

§ 4.8 多变量比估计与回归估计	145
4.8.1 多变量比估计	145
4.8.2 多变量回归估计	147
4.8.3 数值例子——农作物估产调查	147
§ 4.9 二相抽样中的比估计与回归估计	151
4.9.1 比估计	151
4.9.2 回归估计	152
第5章 不等概率抽样	158
§ 5.1 一般描述	158
5.1.1 不等概率抽样的必要性	158
5.1.2 不等概率抽样的分类	154
§ 5.2 放回不等概率抽样与 Hansen-Hurwitz 估计量	155
5.2.1 多项抽样、PPS 抽样及其实施方式	155
5.2.2 Hansen-Hurwitz 估计量及其性质	157
5.2.3 数值例子——职工人数的调查	159
§ 5.3 不放回不等概率抽样与 Horvitz-Thompson 估计量	161
5.3.1 不放回不等概率抽样与包含概率	161
5.3.2 Horvitz-Thompson 估计量及其性质	162
§ 5.4 几种严格的不放回 π PS 抽样方法	164
5.4.1 $n=2$ 的情形	164
5.4.2 $n>2$ 的情形	167
§ 5.5 其他不放回抽样方法及其相应的估计量	169
5.5.1 Yates-Grundy 逐个抽取法及 Das-Raj-Murthy 估计量	169
5.5.2 Rao-Hartley-Cochran 方法及其估计量	173
5.5.3 Poisson 抽样	174
5.5.4 配置抽样	176
5.5.5 不同抽样或估计方法性质的比较	177
第6章 整群抽样	179
§ 6.1 引言	179
6.1.1 定义	179

6.1.2	适用场合及实施理由	179
6.1.3	群划分的原则	180
§ 6.2	群大小相等的情形	181
6.2.1	记号	181
6.2.2	总体与样本平方和的分解	182
6.2.3	群内相关 ρ_c	183
6.2.4	估计量及其方差	187
6.2.5	设计效应	187
§ 6.3	对比例估计的整群抽样	188
6.3.1	群大小相等的情形	189
6.3.2	群大小不相等的情形 比估计	189
§ 6.4	群大小不等的一般情形	191
6.4.1	按简单随机抽样抽群——简单估计	191
6.4.2	按简单随机抽样抽群——比估计	192
6.4.3	对群进行不等概率抽样	194
6.4.4	数值例子——对交通运输量的调查	196
第7章	二阶与多阶抽样	199
§ 7.1	引言	199
7.1.1	定义及适用场合	199
7.1.2	实施方法及同其他抽样方法的关系	200
§ 7.2	二阶抽样——初级单元大小相等的情形	201
7.2.1	记号	202
7.2.2	估计量及其方差	203
7.2.3	最优抽样比例	205
7.2.4	数值例子——生猪存栏量的调查	206
7.2.5	关于比例的估计	208
7.2.6	分层二阶抽样	209
§ 7.3	二阶抽样——初级单元大小不等情形($n=1$)	211
7.3.1	一般说明与记号	211
7.3.2	等概率抽取初级单元	212
7.3.3	不等概率抽取初级单元	213
§ 7.4	二阶抽样——初级单元大小不等的一般情形($n>1$)	215

7.4.1	按多项抽样抽取初级单元	218
7.4.2	不放回抽样时的一般结果	220
7.4.3	按简单随机抽样抽取初级单元	223
7.4.4	按不放回不等概率抽取初级单元	224
§ 7.5	二阶及多阶抽样	227
7.5.1	各级单元大小相等时的三阶抽样	227
7.5.2	多阶抽样中不等概率抽样的应用	229
第 8 章	系统抽样	232
§ 8.1	一般描述	232
8.1.1	定义及实施方法	232
8.1.2	系统抽样与整群抽样和分层抽样的关系	233
8.1.3	系统抽样的优缺点	234
§ 8.2	等概率系统抽样(等距抽样)	235
8.2.1	估计量	235
8.2.2	估计量的方差——用样本群内方差表示	235
8.2.3	估计量的方差——用样本群内相关表示	236
8.2.4	数值例子	237
§ 8.3	方差与总体单元排列顺序的关系	239
8.3.1	随机排列	239
8.3.2	线性趋势	242
8.3.3	单元指标呈周期性变化的情形	243
8.3.4	单元指标呈自相关的情形	243
§ 8.4	具有线性趋势的总体的估计量与抽样方法的改进	245
8.4.1	首尾校正法	245
8.4.2	中位样本法	246
8.4.3	对称(平衡)系统抽样法	246
8.4.4	回归估计量的应用	247
8.4.5	数值例子——部门职工总人数的估计	248
§ 8.5	不等概率系统抽样	250
8.5.1	概述及实施方法	250
8.5.2	估计量	252
§ 8.6	系统抽样中的方差估计	252

8.6.1	等概率系统抽样情形	252
8.6.2	不等概率系统抽样情形	258
第 9 章	复杂样本方差估计的一般方法	260
§ 9.1	引言	260
9.1.1	复杂样本调查	260
9.1.2	方法概述	261
§ 9.2	随机组方法	262
9.2.1	基本思想与方法	262
9.2.2	独立随机组	263
9.2.3	非独立随机组	266
9.2.4	随机组数 k 的选择	267
§ 9.3	Jackknife 方法与 Bootstrap 方法	270
9.3.1	Jackknife 的基本思想与方法	270
9.3.2	有限总体的 Jackknife 方差估计	271
9.3.3	弃 d Jackknife 方差估计	274
9.3.4	Bootstrap 与方差估计	281
9.3.5	d, r 的选取及模拟次数 B 的确定	282
§ 9.4	半样本方法	286
9.4.1	基本思想与方法	286
9.4.2	半样本方差估计性质	288
9.4.3	平衡半样本估计	289
9.4.4	每层多于两个样本单元情况	294
9.4.5	部分平衡半样本估计	295
§ 9.5	Taylor 级数法	296
9.5.1	估计量方差的线性近似估计	297
9.5.2	应用 Taylor 级数于特殊的估计量	299
9.5.3	鞍点逼近方法	301
第 10 章	非抽样误差及相关问题	304
§ 10.1	无回答及其影响	305
10.1.1	无回答的类型	305
10.1.2	无回答的影响	305

10.1.3	多次访问及其模型	306
10.1.4	校正无回答误差的方法	309
§ 10.2	调查误差	311
10.2.1	调查误差的数学模型	311
10.2.2	几种处理方法	315
10.2.3	数值异常情况	318
§ 10.3	敏感性问题的调查	319
10.3.1	敏感性问题的调查与随机化回答	319
10.3.2	Simmons 问题	320
10.3.3	数值例子	321
10.3.4	具多项选择的敏感性问题的调查	322
第 11 章	案例分析	325
§ 11.1	引言	325
§ 11.2	1991 年中国 5 岁以下儿童死亡抽样调查	326
§ 11.3	全国办公自动化设备抽样调查	330
§ 11.4	全国粮食农药污染调查	334
§ 11.5	1987 年中国儿童情况抽样调查	341
§ 11.6	北京地区专业技术人员现状抽样调查	351
§ 11.7	中国 1986 年 74 城镇人口迁移抽样调查	362
§ 11.8	中国妇女社会地位调查	370
§ 11.9	国家卫生服务总调查	385
§ 11.10	人口变动情况抽样调查	397
§ 11.11	农村抽样调查网与抽选方案	402
§ 11.12	人体测量抽样方案	408
附表	随机数表	417
参考文献	422

第 1 章

引 论

§ 1.1 抽样调查的意义和作用

要了解一个国家或地区的人口、环境、资源、社会经济、政治现状,以至人们的意向及对各种问题所持的态度,都必须进行调查。根据调查结果,经过恰当的分析研究,可作为有关领导或决策部门制定政策或采取必要行动的依据。

调查有多种形式,其中最基本的有全面调查、典型调查和抽样调查三类。我国以往在社会经济领域中普遍实行统计报表制度,由统计部门定期将各种统计项目逐级汇总上报,其中大多数项目属于全面调查的范畴。有时,有关部门就一个特定的问题组织大规模的普查,例如人口普查、工业普查、科技普查等都是全面调查。全面调查可以使人们对调查的对象有全面的了解。如果对每个调查对象的调查结果(或他们所提供的资料)都确实无误,且在实际调查过程中,调查对象既没有遗漏也没有重复,数据在各级汇总中也未出现任何差错,那么由全面调查所得的最后结果则是精确而可靠的。但是全面调查也有其本身的局限性。首先,它需要耗费大量的人力、物力和财力;其次,调查所花费的时间也较长。因此对于那些时效性较强的项目,通过全面调查所获得的结果有可能已是过时的信息,从而不能成为决策所需的适时反馈。另外,当调查的对象是无限(或数量极大)时或调查所用的测试方法带有破坏性时,就根本不能采用全面调查。即使在理论上可行,但在实际上由于受到人力、费用与时间上的限制,而不能或不需要进行全面调查。

与全面调查不同的是非全面调查。有许多非全面调查的方法。最重要的是典型调查与抽样调查两种。为某种目的,由调查者选取他认为有“典型”意义的对象进行的调查称为典型调查。例如毛泽东在第一次国内革命战争期间对湖南农民运动的考察与费孝通在 80 年代初期对苏南小城镇进行的社会调查都堪称典型调查的典范。典型调查针对性强,对掌

握事物发展的规律及动向,制定提出相关政策有极大的指导意义。典型调查的主要局限性在于它的调查结果取决于调查的对象,即“典型”的选取以及调查者本人对问题的主观认识。对于那种出于有意或无意的,从并未反映总体情况的“典型”所得的调查结果就容易造成认识的偏差,从而导致决策的失误。至于为了验证自己的某个论点而有意地选择所需的“典型”或事例进行的调查,则更无科学性可言了。典型调查的另一个缺点是由于这类调查通常规模较小,故一般只有定性意义而得不到有关总体的定量结果。

另一种重要的非全面调查方法即是抽样调查(sampling survey)。抽样调查是按照一定的程序,从全体调查对象(我们称之为总体,参见下节)中抽取一部分(称为样本)进行调查,然后根据样本数据对总体目标量进行估计。抽样调查也是一种统计调查方法,它有花费少,适时性强两个基本特点。因此它能以较小的代价及时地获得所需要的信息,这是全面调查所不能比拟的,也是显而易见的优点。但由于抽样调查只对调查对象中的一部分(仅限于所抽到的样本)进行调查,据此对总体进行估计,必然存在误差,即所谓抽样误差(sampling error)。不过这个误差是可以得到控制的,只要抽样足够多,就可使抽样误差任意小。而且对多种抽样方法,可用具体的数量表示抽样误差。事实上,一个经科学设计和严格实施的抽样调查,有可能获得比全面调查更为可靠、更为精确的结果。一项调查的质量不仅取决于调查的规模,更取决于所得数据的正确程度。一个不正确的数据比没有更糟。在抽样调查情形,一则由于调查涉及面较小,参加调查的人员可经过较为严格的统一培训,其素质和经验都可比参加同样性质的全面调查(普查)的工作人员要高。另外,在抽样调查中,更有可能采用精确和可靠的分析测试手段。因此抽样调查获得的原始数据一般比全面调查所获得的相应数据更为精确。另一方面,也由于总的工作量相对较小,抽样调查的全过程可以通过各种措施(这些措施比全面调查时采用的类似措施容易实行得多)使整个调查过程处于控制状态。最后,也是相当重要的一点是:在社会经济和其他某些领域的一些调查,若采用全面调查,被调查的单位或个人容易将调查与对单位或本人的评价联系起来,或认为调查结果直接与本单位或本人的某种利益有关,从而会发生不能如实填报或回答,甚至人为干预等情况。例如我国历年耕地面积数字按报表汇总的数字就比实际数字偏小;公安机关掌握的一个地区的出生率和婴儿死亡率都可能比实际数字偏低;农产量在一个时期可能有

虚报的倾向,而在另外一个时期,又可能有瞒产的倾向等等。所有这些人 为的偏差大大影响了调查结果的正确性。而抽样调查由于不涉及每个单位或每个人,一般地说,没有单位之间或个人之间比较的意义,从而能在相当程度上减轻被调查单位或个人的心理压力,较为愿意提供真实数据。从以上几个方面看,抽样调查完全有可能做到比全面调查更为精确和可靠,再加上它的经济与快速,从而乐于被人们所采用。它的应用也就愈来愈广泛了。

当然,抽样调查并不能完全取代其他调查方法。当我们需要弄清某些社会现象的机制或发展趋势时,仍需要进行典型调查。而全面调查过去是,今后也将继续是我国统计部门和其他一些部门的一种基本的调查手段。但是,抽样调查作为一种科学的调查方法,其重要性必将日益显示出来。它不仅可以在一些项目中单独使用,而且也可以与全面调查或典型调查结合起来,起到相互补充的作用。例如在人口统计中,我国今后将每 10 年进行一次普查,而在其他年份进行人口变动情况的抽样调查,以此来估计每年的人口数。即使在普查时,也常同时采用抽样调查的方法对普查结果进行核对和进行修正。反之,以前的普查资料也为抽样调查提供了丰富而可靠的背景材料,从而能使抽样调查获得更好的效果。

§ 1.2 若干基本概念

从方法论意义而言,抽样调查属于应用统计。为了便于在以后各章讨论具体抽样方法,在这一节中我们解释抽样调查中的一些基本概念,将抽样调查中的某些问题用一般的统计语言进行描述,并指出它们与数理统计其他分支中相应问题的一些区别。

1.2.1 总体与样本

总体与样本是统计中最基本的概念。总体(population)就是所研究(调查)对象的全体。例如在全国儿童情况调查中,全国所有 0~14 岁的儿童就构成调查的总体。调查的目的是为了得到有关这个总体的某些参数,例如全国儿童总数,每个年龄组男女儿童的平均体重,学龄儿童的在校率等等的估计。因此调查时必须涉及有关指标(characteristic)。总体是由个体(item, individual)组成的。作抽样调查时,我们按某种方法从总体中只抽取其中一部分个体进行调查。这部分个体就称为样本

(sample). 在儿童调查中, 全国每个0~14岁的儿童就是一个个体, 而根据设计的抽样方案抽到的需要进行调查的儿童构成样本. 根据被抽到的这些儿童的各调查指标的数据(即样本数据), 即可对总体参数或调查的总体目标量进行推算即估计. 用样本推断总体是数理统计最基本的特征.

在理论上以及实际处理时, 抽样调查中的总体通常假定是有限的. 尽管在实际问题中, 不乏存在总体很大甚至无限的情形, 但通过划分抽样单元(详见1.2.3段)的方法, 总体就可以看作是有限的, 而且在一般情形, 总体大小 N 已知. 这一点与数理统计中通常讨论的无限总体是有区别的. 另外, 抽样调查中的个体(或抽样单元)都是具体的, 且是可以辨别的, 因此相应的总体也是具体的. 这又与统计中许多场合不同, 例如试验设计中的总体就只是一个抽象的总体. 在讨论抽样调查具体方法时, 一般很少对总体进行什么假定, 特别是很少对它的分布作任何假定, 这也是因为它过于具体的缘故.

抽样调查所处理的样本一般比较复杂. 在绝大多数情形, 样本中的观测数据不是独立同分布的, 这也与数理统计中通常讨论的情形不同. 因此有人将抽样调查中获得的数据称为“不干净数据”, 称这样的样本为复杂样本(complex sample).

1.2.2 概率抽样

抽样调查中的一个基本问题是样本的抽取方法, 也即抽样方法. 抽样又可分为概率抽样(probability sampling)和非概率抽样(non-probability sampling)两类. 概率抽样也称随机抽样(random sampling). 但当使用后一个术语时, 要注意它与另一个术语即随机抽取(to draw an item at random)的区别. 从由 N 个个体组成的总体中抽取一个个体时, 若总体中的每一个体被抽到的可能性都相等, 则称这种抽取方法为随机抽取. 因此, 随机抽取是指等概率从总体中抽取个体的方法. 而概率抽样的含义比它更为广泛, 它是一种从总体中按一定概率获取样本(一组个体)的方法. 概率抽样具有如下基本特点:

- 1) 能够确切地定义(或区分)不同的样本, 即能够明确表明一个确定的样本包含哪些个体;
- 2) 对每个可能的样本, 都赋予一个被抽到的概率;
- 3) 通过某种随机形式从总体中抽取一个样本, 使这个样本被抽中的概率等于所赋予的概率;

4) 从样本估计总体参数时需与抽样概率相联系。

在实际问题中, 抽样可以逐个进行, 即每次只从总体中抽取一个个体(或单元), 也可以整个样本一次同时抽取。在逐个抽取时, 每次被抽到的个体可以不放回也可以重新放回总体中去。前者称为不放回抽样(sampling without replacement), 后者称为放回抽样(sampling with replacement)。如果整个样本一次同时抽取也是一种不放回抽样。另外, 当抽取总体中的每个个体(或尚未进入样本的个体)时, 个体被抽中的概率可以是等概率的, 也可以是不等概率的, 前者称为等概率抽样(sampling with equal probabilities), 后者称为不等概率抽样(sampling with unequal probabilities)。

概率抽样的优点是能够保证样本的代表性, 避免人为的干扰和偏差。它还能对由于抽样引起的误差——抽样误差进行估计。因此采用概率抽样可以获得估计的精度。鉴于这两个原因, 概率抽样是最科学、应用最广泛的一种抽样方法。因此只要有可能, 就应尽量采用概率抽样。

有时概率抽样在实际中难以实现, 例如从一间货物堆得很满的仓库中进行抽样, 或从大气或江河海洋中采取大气样或水样, 这时样本通常只能在局限于总体的某一部分中抽取, 而且也难于严格地按一定的概率原则来进行抽样。也有这种情况, 由于经费和时间的限制而不能进行严格的概率抽样。在这些情形, 就只能采用某种非概率抽样。一种常用的非概率抽样是所谓的判断抽样, 或称经验抽样。这种抽样是根据抽样者的主观经验和判断, 从总体中选择“平均”的或认为有代表性的同时又容易取得的个体作为样本。当总体变差较大, 而抽样的数量又不能很大时, 判断抽样有可能提供比概率抽样更为准确的估计。这是因为判断抽样的精度主要取决于抽样者的经验, 与样本量(sample size)关系不大; 而概率抽样的精度主要取决于样本量。除了有主观随意性外, 判断抽样的另一个缺点是不可能定量获得估计的精度。还有一种常见的非概率抽样形式, 此时样本完全或几乎完全由“志愿者”所组成, 例如刊登在报刊杂志上的读者意见的调查, 调查表是否寄回完全随读者的意愿决定, 因而这种调查结果只是反映了这部分“热心”读者的意向。在某种意义上看, 这根本就不能算是一种抽样调查了。

1.2.3 抽样单元与抽样框

为使概率抽样能够实施, 同时也为了具体抽样的便利, 通常将总体划

分成互不重叠且又穷尽的若干部分, 每个部分称为一个抽样单元(sampling unit)。抽样单元不一定是组成总体的最小单位, 即前面所说的个体, 但有时候也可直接把个体作为抽样单元。总体中的抽样单元数一定是有限的, 而且是已知的。这正是在1.2.1段中提到的, 我们总是把总体局限为有限总体的缘故。抽样单元的划分可以有较大的选择余地, 例如在电视收视率抽样调查中, 可以将每个电视观众作为抽样单元, 也可以将每个拥有电视机的家庭作为抽样单元; 在人口变动量抽样调查中可以将县、乡(街道)或居民委员会(村)作为抽样单元。抽样单元可以是自然形成的, 例如各级行政单位、机关、学校、工厂以至个人; 也可以是人为划分的, 例如为调查田地中的害虫总数, 将整块田块划分成每边长一米的正方形小块, 而将每个小块作为一个抽样单元。抽样单元又可以有大小之分, 一个大的抽样单元(例如省)可以分成若干个较小的单元(例如县)。前者称为初级单元或一级单元(primary sampling unit, 简记为PSU), 后者称为次级单元或二级单元(secondary sampling unit)。次级单元又可分为更小的三级单元、四级单元等。将抽样单元分级, 主要是基于具体抽样方法的考虑, 例如多阶抽样与整群抽样。

在总体中按抽样单元进行概率抽样时, 需要一份有关抽样单元的名册、清单或地图。记录或表明总体所含全部(初级)抽样单元或一个较大抽样单元所包含全部次一级抽样单元的这种名册、清单或地图称为抽样框(sampling frame)。在抽样框中, 每个抽样单元都被编上号。抽样框是设计并实施一个抽样方案所必须具备的基础资料。一旦某个单元被抽中, 也需要根据抽样框在实际中找到这个单元, 从而能够实施调查。

1.2.4 总体参数的分类

抽样调查的主要目的是通过样本对我们感兴趣的某些总体参数进行估计, 这些总体参数也就是调查的目标量。通常需估计的总体参数可以归纳为以下几类:

1) 总体总和(population total): 例如全国人口数, 一个地区某年的粮食总产量、我国大熊猫的现存数量等。

2) 总体均值(population mean): 例如职工平均月工资、粮食中平均残留的“六六六”农药的含量、某地区粮食的亩产量等。

3) 总体中具有某种特定特征的个体总数或它们在总体中所占的比例或百分率(proportion or percentage): 例如某地区人口中在上一年度

死亡人数或死亡率、育龄妇女生育率、结核病患率等。

4) 总体两个不同指标的总和或均值的比值(ratio): 例如家庭中用于食品的支出在总支出中所占的比例、某地区学龄儿童的在校率(若该地区学龄儿童总数也要通过调查才能估计)等。

5) 总体分位数: 例如我国成年人身高、胸围、腰围等人体尺寸的5%, 50%, 95% 分位数等。

上述五种总体参数中的前四种都有不同程度的内在联系。若记总体中第 i 个单元的某个调查指标值为 $Y_i (i=1, 2, \dots, N)$, 则总体总和

$$Y = \sum_{i=1}^N Y_i \quad (1.1)$$

与总体均值

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i = \frac{Y}{N} \quad (1.2)$$

只相差一个已知的常数 N 。而若令

$$Y_i = \begin{cases} 1, & \text{若总体中第 } i \text{ 个单元具有所考虑的特征;} \\ 0, & \text{否则。} \end{cases} \quad (1.3)$$

则总体中具有这种特征的单元总数 $A=Y$, 比例 $P=Y$ 。因此前三种总体参数在数学处理意义下是等价的。

至于总体两个不同指标的总和或均值之比

$$R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}} \quad (1.4)$$

有别于3)中的 P , 因为此时 X (或 \bar{X}) 也需要从样本中估计。因此 R 与 P 的处理, 特别是对它们估计量的精度进行估计时要采用不同的处理方式。

1.2.5 误差来源与精度表示

抽样调查中的误差来源主要有两个。一种称为非抽样误差(non-sampling error), 它是由于调查中获得的原始数据不正确(例如测量误差)、抽样框有缺陷(抽样框中的抽样单元有重复或遗漏)、或在调查中由于种种原因无法得到按抽样设计方案的全部样本数据(例如部分调查对象拒绝回答问题, 等原因引起的。这种误差在全面调查中也是普遍存在的。为减少非抽样误差, 必须通过改进调查表的设计或测试方式, 严密调查组织, 提高调查员的素质以及加强调查中各个环节的质量控制, 才能见效。对某些问题, 例如测量误差以及不回答误差(non response error)

对调查结果的影响需根据具体情况特殊处理。另外,对于不易获得被调查者真实情况的诸如敏感性问题的调查也必须通过设计特殊的调查方法进行处理。

抽样调查误差的另一来源是由于我们实际上是用局部的样本数据对整体的总体参数作出估计所引起的误差,这部分误差称为抽样误差。抽样误差愈小,估计量的精度就愈高。在本书中主要考虑这种误差。

若令 $\hat{\theta}$ 是通过样本获得的对总体某个参数 θ 的估计,则抽样误差一般用以下的均方误差(mean square error)来表示:

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2, \quad (1.5)$$

式中的 E 表示数学期望(均值)。由于 θ 是未知的,因此均方误差并不总是能够得到的或精确估计的。均方误差可以分解成两个部分:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2. \end{aligned} \quad (1.6)$$

上式中的第一项

$$V(\hat{\theta}) \triangleq E[\hat{\theta} - E(\hat{\theta})]^2 \quad (1.7)$$

是 $\hat{\theta}$ 的方差(variance),而第二项

$$B^2(\hat{\theta}) \triangleq [E(\hat{\theta}) - \theta]^2 \quad (1.8)$$

是 $\hat{\theta}$ 的偏倚(bias) $|E(\hat{\theta}) - \theta|$ 的平方,偏倚为零的估计量,也即满足

$$E(\hat{\theta}) = \theta \quad (1.9)$$

的估计量 $\hat{\theta}$,称为无偏估计量(unbiased estimator)。对于无偏估计量,它的均方误差即是它的方差。

有时也用相对均方误差(relative mean square error) $\text{MSE}(\hat{\theta})/\theta^2$ 或相对方差(relative variance) $V(\hat{\theta})/\theta^2$ 来表示 $\hat{\theta}$ 的精度。

如果一个估计量的偏倚及方差都随着样本量 n 的增大而减小,而且偏倚比均方误差的平方根小得更快,即

$$\lim_{n \rightarrow \infty} \frac{|E(\hat{\theta}) - \theta|}{\sqrt{\text{MSE}(\hat{\theta})}} = 0. \quad (1.10)$$

则称这个估计量是可用的(feasible)。对可用估计量,只要 n 足够大, $\hat{\theta}$ 的精度主要取决于 $\hat{\theta}$ 的方差,或它的平方根 $S(\hat{\theta}) = \sqrt{V(\hat{\theta})}$,即 $\hat{\theta}$ 的标准差(standard deviation)。

由于我们通常不对总体分布作任何假定,又因为样本的复杂性,在抽样调查中,一个估计量 $\hat{\theta}$ 的精确分布是无法求得的。但近期的研究表明,在某些假定下,一定类型的复杂样本(例如分层随机样本)估计量的分

布, 在大样本时是近似正态的. 在一般情形, 虽然没有严格的理论证明, 但许多模拟结果也得出类似结论. 据此, 对于一个可用的估计量 $\hat{\theta}$, 只要样本量 n 足够大, 可以构造 θ 的给定置信水平 $1-\alpha$ 的近似置信区间:

$$\hat{\theta} \pm u_{\alpha} \sqrt{V(\hat{\theta})} \quad \text{或} \quad \hat{\theta} \pm u_{\alpha} S(\hat{\theta}), \quad (1.11)$$

其中 u_{α} 是标准正态分布的双侧 α 分位数. 例如若取 $\alpha = 0.05$, 则当 n 大时, $\hat{\theta}$ 的置信水平为 95% 的近似置信区间为:

$$[\hat{\theta} - 1.96 S(\hat{\theta}), \hat{\theta} + 1.96 S(\hat{\theta})]. \quad (1.12)$$

(1.11)式中的

$$d \triangleq u_{\alpha} S(\hat{\theta}) \quad (1.13)$$

称为 θ 的 $1-\alpha$ 置信水平下的绝对误差限, 它满足

$$P_r(|\hat{\theta} - \theta| \leq d) = 1 - \alpha. \quad (1.14)$$

同样可以定义相对误差限 r , 它满足

$$P_r\left(\left|\frac{\hat{\theta} - \theta}{\theta}\right| \leq r\right) = 1 - \alpha. \quad (1.15)$$

r 可由下式确定:

$$r = u_{\alpha} \text{Cv}(\hat{\theta}) = u_{\alpha} S(\hat{\theta}) / \theta. \quad (1.16)$$

其中 $\text{Cv}(\hat{\theta})$ 是 $\hat{\theta}$ 的变异系数 (coefficient of variation).

1.2.6 样本量、费用与效率

样本量即是样本中包含的抽样单元的数目. 通常为了便于比较起见, 它是以最小抽样单元或个体为单位计算的. 样本量的确定也是抽样调查中的一个重要问题. 样本量愈大, 抽样误差就愈小, 估计量的精度就愈高. 但样本量又直接与费用有关. 样本量愈大, 调查的费用也就愈高. 最简单的费用函数是如下的线性费用函数, 总费用

$$O = c_0 + cn, \quad (1.17)$$

其中 c_0 是与样本量 n 无关的固定费用, 包括组织、宣传、抽样框的准备等; 而 c 是平均每抽一个单元的费用, 包括调查本身的费用、旅费以及数据处理费用等.

因此, 一个好的抽样设计必须同时考虑精度与费用两个因素. 对于一个具体的抽样设计, 应尽量做到在固定的费用限制下使精度最高, 或在要求达到的精度条件下, 使调查的总费用最省. 换言之, 我们要求设计的效率最高, 这样的抽样设计称为最优抽样设计.

§ 1.3 几种基本的抽样方法

对不同项目应采用不同的抽样方法。最基本的抽样方法有以下五种。在实际问题中,一个具体的抽样方案大多是这五种方法的各种形式的组合。

1.3.1 简单随机抽样 (simple random sampling)

简单随机抽样也称为单纯随机抽样。从大小为 N 的总体中逐个不放回地抽取 n 个单元组成样本,每次抽取对当时尚未入样的单元都是随机抽取的,也即都是等概率的。简单随机样本也可从总体中一次取得全部 n 个单元,只要全部可能的 $\binom{N}{n}$ 种这样的样本每种被抽得的概率都相等(都等于 $1/\binom{N}{n}$)。注意,这里所用的“简单随机”与一般数理统计文献中的含义不同。一般数理统计书中所谓的简单随机样本是指在无限总体中独立抽样所得的样本或在有限总体中放回随机抽样抽得的样本,因而是独立同分布样本 (independently identically distributed sample),但在本书中保留绝大多数有关抽样调查文献中对简单随机抽样的定义。

简单随机抽样是所有其他抽样方法的基础。因为在理论上最易于处理。这种方法表面上看简单易行,但在许多实际情形实施时有很大的困难。主要原因是它需要一个对全部基本单元的完整抽样框,且所得的样本单元相当分散,调查不便。因此在大规模抽样调查中,很少单独采用简单随机抽样。尽管如此,它依然是所有其他抽样方法的基础。

简单随机抽样中的估计方法,通常是用样本平均数来估计总体均值,这就是所谓简单估计。在有辅助变量可以利用时,为提高估计精度,也可以用比估计和回归估计等方法。

1.3.2 分层抽样 (stratified sampling)

将总体中的单元按某种原则进行划分成为若干个子总体,每个子总体称为层。在每层中独立进行简单随机抽样或其他抽样,这样的抽样就称为分层抽样。分层抽样的估计先对各层进行,然后再综合对总体参数进行估计。

分层抽样适用于调查本身既需要对总体进行估计,也需要对局部

(层)进行估计的情况。分层抽样实施和组织都比较方便。当层内单元指标差异较小,而层间单元指标差异较大时,采用分层抽样可以大大提高估计的精度。例如在家庭调查中,将住户家庭按城市、农村以及不同职业分层,由于不同层家庭的收支水平和生活习惯相差较大,因而这样的分层抽样精度较高。

1.3.3 整群抽样(cluster sampling)

若总体中的每个抽样单元可以分成若干次级单元,抽样仅对初级单元抽,若某个初级单元被抽中,则调查这个单元中所有次级单元,这种抽样方法称为整群抽样。这里的群(cluster)就是指初级单元。例如为对我国成年人的身体尺寸进行调查,确定对每个人要测量 76 项指标,为此需组织专业测量队。如果被测量的人相对集中,显然就比较方便,可以大大节省调查费用。这是实施整群抽样的主要考虑。因此在这种调查中,我们对单位抽样,然后测量被抽中单位的每一个职工。整群抽样的缺点是效率不够高。由于一个群内的(次级)单元多少有点相似,故对每个次级单元都进行调查会造成浪费。故若按总样本量(按小单元计算),整群抽样的精度比直接对总体中所有次级单元进行简单随机抽样低,但这可以通过适当地多抽样来得到弥补,从总体上有可能在总费用相同的条件下获得更高的精度。

1.3.4 多阶抽样(multi stage sampling, subsampling)

多阶抽样也称多级抽样。若初级单元内的次级单元相似程度较大,正如前面所说的那样,调查所有次级单元会造成很大的浪费。此时一个自然的想法是在被抽中的初级单元中再对次级单元进行抽样,这就是二阶抽样。类似的,可用三阶抽样、四阶抽样等。例如在全国抽省、省中抽市、县,市、县中抽区、乡或镇等等。多阶抽样既保持了样本相对集中,又避免了不必要的浪费,而且实施也比较方便。它也不需要每个初级(或二级)单元都有一个完整的抽样框。但多阶抽样的估计比较困难。

1.3.5 系统抽样(systematic sampling)

若总体中的抽样单元按某种次序排列,在规定的范围内随机抽取一个(或一组)初始单元,然后按一套事先确定的规则确定其他样本单元的抽样方法称为系统抽样。与其他几种抽样不同的是:这里只有初始单元

是经随机抽取的,其他样本单元都随着初始单元的确定而确定。最简单的系统抽样是在取得一个初始单元后,按相等的间隔抽取其他样本单元,这就是所谓的等距抽样。系统抽样的主要优点是实施方便,不需对所有样本单元进行随机抽取,也不一定需要一个完整的抽样框。如果对总体单元的指标按其排列次序的变化规律有所了解,并加以合理利用的话,系统抽样的效果也很好。它的主要缺点是,在多数情形得不到估计量的简单的精度估计。事实上,许多实用而行之有效的系统抽样并不属于严格的概率抽样。

上面对几种常用的抽样方法作了简单的介绍,在实际运用中会有许多变化。例如在某些方法中,入样单元既可放回也可不放回;可以进行等概率抽样,也可进行不等概率抽样。在具体设计抽样方案时,还要考虑多种复杂的因素。这些将在以后各章中分别详述,并在最后一章实例分析中进行讨论。

§ 1.4 抽样调查的步骤

对抽样调查,不同的项目所包含的步骤也不尽相同,但大致上都包含以下几个重要的步骤:

1) 明确调查的目的,确定调查方式和所需估计的目标量:通过一次调查要达到什么目的?调查哪些指标?需要估计哪些目标量?都是首先需要明确的。因为调查的具体形式和组织,抽样方案的制定以及调查数据的处理都取决于调查的目的和调查的目标量。抽样方案确定后,调查目标量的任何改动,往往会使已制定的方案不再适用。因此在此阶段,主持单位必须会同有关专家进行反复讨论和审定。在这一步骤中,首先要确定总体范围及抽样单元。这个问题有时并不简单,例如对残疾人的调查,首先要明确残疾人的划分标准。在确定需要估计的目标量也即总体参数时,要注意防止列入过多的调查项目(指标)。项目过多,不仅会增加调查和以后数据处理的费用和时间,还可能使不回答率增加,并影响原始数据的质量,因而是得不偿失的。在这一阶段还必须同时确定调查的方式,是采用当面询问还是通过调查表(或称问卷, questionnaire)或者是两者结合?对于需要技术测试的调查,还要确定测试或分析的方法。对调查的目标量应提出具体的精度要求,作出调查的经费预算,确定调查的标准时刻等。对调查表,应在认真设计的基础上,征求有关专家的意见,反复

修改,力求完善。在许多情形,在正式调查前往往还需进行一次试调查(pilot survey)。

2) 抽样设计,给出相应的数据处理公式:这是一个抽样调查中总体设计的最重要部分,包括选择抽样方案的类型,确定样本的抽取方法及样本量。在制定具体的抽样方案时,既要考虑方法的科学性又要照顾到实际的可行性。例如设计一个全国性抽样调查,需要考虑一个多阶抽样,此时前一、两阶抽样是关键的,必须采用一些效率高的抽样方法,复杂一些也无妨,因为前一两阶的抽样可以由设计者自己来实施。与此相反,对最后一、两阶抽样则由于涉及基层就必须采用尽可能简单的抽样方法。在制定抽样方案时,还必须同时考虑到调查以后数据处理问题,给出与抽样设计相匹配的总体参数的估计公式以及估计量的精度公式。

在抽样方案确定以后,就要实施抽样,即确定需要调查的样本单元,为此,在事先要准备好相应的抽样框及其他有关资料。

3) 调查的实施,即取得样本数据的过程:为保证调查的质量,确保原始数据尽可能正确,应建立相应的职能办公室,事先进行调查员的培训,制定并采取各种质量控制措施等。

4) 数据处理:现代大规模的抽样调查所获得的数据,一般都在电脑上进行处理。首先是编码并录入数据,建立数据库。在正式处理前,要对已录入的数据进行编辑加工,按一定规则检出并处理原始数据中存在的或在录入过程中混入的异常数据。检查方法应同时采用统计检查及逻辑检查。经过反复检查的数据即可进行进一步的处理。其中最重要的是按在抽样设计时给出的总体参数的估计公式与估计量精度公式,计算每个目标量的估计值及其相应精度,特别是方差与变异系数的估计值。有时还需要结合分析目的进行其他统计处理,例如列联表分析与多元分析等。

5) 调查结果的分析,提出最终报告:根据数据处理的结果以及调查目的,对调查结果进行综合分析,提出最终的调查报告。

抽样调查的全过程可用下述的方框图来表示(图 1.1)。

除以上所述的步骤外,实行一项抽样调查还有相当多的行政组织与部门的协调工作。例如,根据我国的统计法,任何一项统计调查,如果在本系统内进行,要在相应的统计部门备案;如果调查对象涉及本系统以外的部门或人员,要向相应的统计部门申请,在得到批准后,方可以进行,否则,调查对象有权不予回答。因此,在进行抽样调查之前要履行必要的手

续,使调查合法并使调查结果受到法律保护。

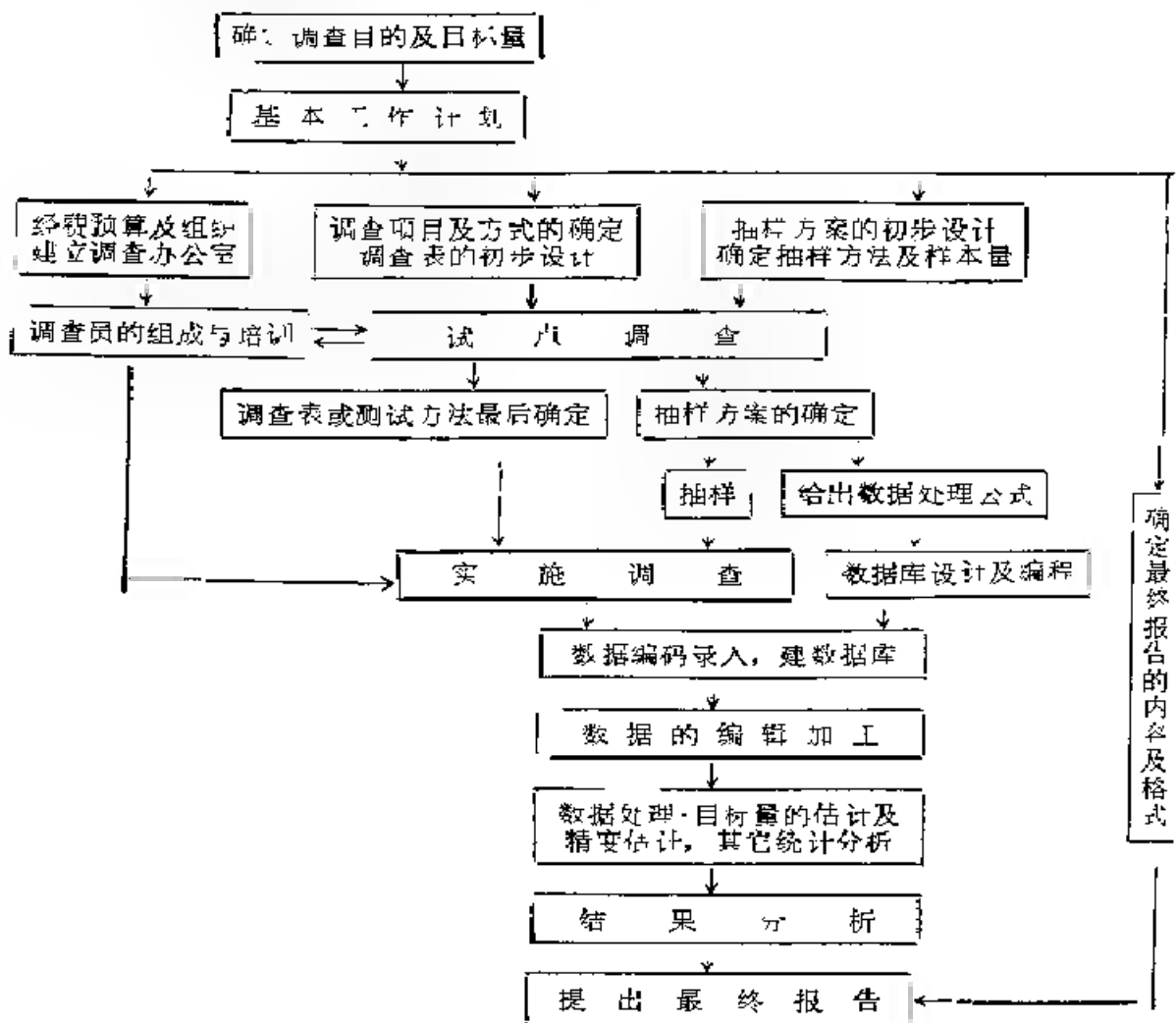


图 1.1 抽样调查流程图

§1.5 抽样调查的应用与历史发展

1.5.1 主要应用领域

抽样调查的应用范围极为广泛。要罗列全部可能的应用领域是不可能的。一般地说,凡需获得与一个较大系统(无论是社会的还是自然的)有关的信息,都可以应用抽样调查的方法。以下是几个主要的应用领域。

1) 人口调查。这是最早应用抽样调查的领域,目的是对一个国家或地区的人口总数、构成及其变动情况作出估计。调查包括妇女生育、儿童出生、人口死亡及迁移等等内容。

2) 经济调查: 包括对各种产业、农业、畜牧业、工商贸易、交通、市场和物价调查等。

3) 社会调查: 包括住户(家计)调查、劳动就业、文化教育、妇女地位、婚姻、儿童情况以及犯罪情况调查等等。

4) 民意调查: 这是一种特殊的社会调查, 目的是了解人们对各种政治、社会、经济等方面问题的态度、要求以及对某项政策或候选人的支持程度等等。

5) 卫生调查: 公共卫生情况、预防接种覆盖情况、疾病流行、病因及治疗后的随访等等。

6) 环境资源调查: 有关耕地、森林、草原、能源、动物、害虫的估计以及对大气、水质等环境污染情况的监测等。

7) 人体测量: 即对人体各部位尺寸的测量调查、用于各种人类工效学标准(包括服装号型的制定)等等。

除此以外, 抽样技术还广泛地用于各种普查数字、统计报表数字可靠性的核对与检查、帐目的审计以及各种工业产品和材料的质量、服务质量的调查评估等等。

1.5.2 国际发展简史

抽样调查是统计中应用与发展最早的一个分支。从部分推算整体的思想由来已久。早在 1662 年英国人 J. Graunt 曾对伦敦城内保有较完整登记表册的教区作家庭调查, 他根据一个教区的洗礼和葬礼次数来估计当时伦敦的总人口约为 384,000 人。17~18 世纪的人口统计学家包括英国的 W. Petty 和 F. Halley, 瑞典的 P. Wargentin 以及德国的 J. P. Susmich 都曾根据一个地区的部分数据资料对整个地区作过类似的推算。更完整的工作要数法国著名数学家 P. S. Laplace, 他在 1786 年写的一篇关于巴黎人口出生、死亡和婚姻状况的论文里, 就建议用某些地区的出生率来推算整个法国的人口, 并对推算出来的结果的误差问题进行了研究。1802 年, 他在法国政府的支持下, 作了一次统计抽样的实验。他在全法国挑选了 30 个社区, 这些社区的选择既要抵消气候差异等地区影响, 同时要求能够提供最精确的信息资料。对这些社区连续三年出生的人数进行分析, 他发现平均每 28.35 个居民中每年出生一个婴儿, 也即出生率为 35.27%。据此他推算出到 1812 年时, 在法兰西帝国疆域内, 每年出生人数为 150 万人, 全国总人口为 4253 万。他甚至还给出了推算

出来的人口数与实际人口数之间的误差为 0.86%。

不过像 Laplace 这样的工作在当时还是比较零星的,较完整的抽样调查工作起始于上世纪 90 年代。那时许多欧洲国家相继在社会经济领域中应用抽样调查。当时担任挪威统计局长的 A. N. Kiaer 在 1891 年利用抽样调查估计挪威全国国民的收入和财产情况,用以研究该国人口的一些经济和社会特征。1901 年丹麦进行了农产品产量的抽样调查。英国的 A. L. Bowley 等人也在 1906 年及 1913 年进行了社会经济方面的抽样调查。在第一次世界大战期间,美国曾用抽样调查制定军服尺寸的系列标准。这些都可以作为抽样调查的一些早期应用。

这些早期的抽样调查,在样本抽取时广泛使用了所谓“代表性调查”(representative investigation)方法。这也是首先由 Kiaer 提出来的,他认为抽样调查的准确性主要不取决于样本大小,而在于样本的代表性。他的思想是使样本成为总体的一个缩影,样本单元不是随便选取的,不应有主观偏误,要求对调查的可靠性进行评价。显然, Kiaer 的这些观点在当时是包含许多合理成分的。故在 1908 年国际统计学会 (International Statistical Institute 即 ISI)通过一项决议,引导和支持采用代表性方法。但当时对如何才能抽到有“代表性”的样本,意见并不统一。1926 年 ISI 指出在选取代表性样本的许多方法中,要区分两种方法,即随机抽取和有目的地或有意地选取样本单元,后者是尽可能使抽到的样本单元合并起来能产生与总体相近的特性。这就是目的性抽样 (Purposive selection), ISI 强调需要随机抽取样本。

随着抽样调查实际的需要以及统计基本理论的发展。从本世纪 20 年代起以及其后的两个年代里,抽样调查的基本理论也就逐步形成了。在这中间首先要提到 Bowley, R. A. Fisher 和 J. Neyman 等人的贡献。上面提到的 ISI 的关于抽样方法的推荐就包含了 Bowley 的许多建议。他提出按抽样框根据随机或系统方法进行抽样以及比例分配的分层抽样方法,他还强调了不回答问题对调查可能产生严重的影响。

统计大师 Fisher 从 1919 年起在英国罗萨姆斯特德 (Rothamsted) 实验站长达十余年的工作期间发展了近代实验设计与方差分析的理论与方法。在实验设计中, Fisher 提出随机化 (randomization)、重复 (replication) 及区组 (block) 三个基本原则。这三个重要原则也同样对抽样的理论发展提供了基础。随机化是获得无偏估计的基础,采用重复技术使得方差估计能够在抽样获得的数据基础上得以进行;而划分区组用

于抽样即是分层，目的是为了减少抽样误差。为了提高效率，Fisher 及其在 Rothemsted 的同事们采用了多阶抽样。Fisher 以后，F. Yates 领导了该实验站的实验统计部门，他对抽样调查的主要贡献在于关于系统抽样的研究以及二次世界大战后受联合国的委托在人口抽样调查方法的研究。

J. Neyman 对抽样调查的贡献是在本世纪 30 年代。他在 1934 年的工作为从有限总体中的抽样奠定了基础。他明确指出：在这以前的任何对目的性抽样给予的理论描述不外是分层和整群随机抽样，从而排除了专门对目的性抽样另作理论探讨的必要性。因此，Neyman 大力提倡随机抽样。他对抽样调查的另一重大贡献是大家熟知的建立了置信区间的理论。抽样方法能否为人们所普遍接受的一个根本问题是能否对抽样误差或精度给予科学的描述。过去有许多人试图借助正态分布去计算抽样误差，但未能对估计值的精度作出有效的解释。Neyman 成功地建立了现在作为经典统计最基本概念之一的置信区间理论。这个理论的产生是在对抽样的代表性这个问题的研究基础上产生的。Neyman 还对分层抽样中的最优分配、整群抽样和比估计理论等作出了重要贡献。其实，关于分层抽样最优分配结果在 Neyman 以前就已由前苏联统计学家 Tschuprow 在 1923 年给出了。只是由于当时苏联与外界隔绝，未被其他学者注意到。其实自 19 世纪下半叶起，俄国在调查统计方面的工作已位于世界前列。十月革命后，列宁领导下的苏维埃政府也对抽样调查相当重视。只是后来苏联当局自批判摩尔根遗传理论，株连统计方法在社会经济上的运用，才使抽样调查与其他数理统计的理论研究在很长一段时期内陷于停顿。

20 年代末与 30 年代初世界经济大萧条产生的新交易计划以及对于经济信息的需求促进了美国在 30 年代起进行无数次大规模的调查。另一个促进抽样调查的应用与研究的因素是对社会舆论调查即民意测验的影响。1935 年美国著名的盖洛普(Gallup)民意调查所成立。翌年，它通过随机抽样的原则对选民进行调查，成功地预测了当年美国总统选举的结果，从而使它的声名大振。总之，30 年代以来，美国逐渐成为抽样调查理论和方法的发展中心。

早在 1933 年，G. W. Snedecor 在美国衣阿华(Iowa)州立大学(位于 Ames)建立了统计实验室。该实验室与农业部及商业部的普查局(the Bureau of Census)进行了一系列合作研究。特别是在多阶抽样设计及

其优化问题,以 PPS 抽样为代表的^①不等概率抽样的引进、系统抽样的理论与经验研究、调查设计中辅助信息的利用、非抽样误差、主样本(master sample)的设计与应用以及地区抽样(area sampling)的研究构成了现代抽样调查理论与方法的重要内容。先后在这些机构工作的有 A. J. King、R. J. Jessen、W. G. Cochran、H. O. Hartley、F. Stephan、M. N. Hansen、W. N. Hurwitz、J. N. K. Rao 与 W. A. Fuller 等著名学者。此外在 Michigan 大学, J. R. Goodman、L. Kish 主持的抽样研究中心在控制抽样(controlled selection)以及改进调查表设计和数据搜集程序方面都作了大量工作。近年来, L. Kish 还领导了在调查方法方面的国际培训。

在其他国家,特别是在第三世界中,应该特别提到的是印度统计学家的贡献。早在本世纪 30 年代, P. C. Mahalanobis 创建了印度统计学院,成为印度抽样调查的权威机构, Mahalanobis 特别注重抽样设计的实效,即在费用与精度之间取得平衡的最优设计。他还最早提出了交叉子样本(interpenetrating subsamples)的概念,后来成为估计复杂样本估计量方差的重要方法——随机分组(random group)及其他重复方法(replicated methods)的基础。另一位印度统计学家 P. V. Suklatac 对分层抽样及非抽样误差估计等方面也作出了重要的贡献。

此外,加拿大学者 N. Keyfitz 与 I. Fellegi 及瑞典学者 T. Dale-nius 等在抽样调查的理论与实践中也作出了重要的贡献。

最后我们应该提及联合国对抽样调查的发展和推广所起的作用。联合国统计司(Statistical Office of the United Nations)早在 1947 年就成立了以 R. A. Fisher 为顾问, P. C. Mahalanobis 为主席的包括 F. Yates、W. E. Deming 等人的抽样分委员会(UN Subcommittee on Sampling)发表了一系列指导性的文件。这些文件包括专题报告和手册,为其成员国,特别是第三世界国家抽样调查的应用与推广,改进这些国家统计数字的质量都起了极大的作用。

1.5.3 我国抽样调查的应用与发展

由于历史原因,我国抽样调查的研究与应用起步较晚。长期以来,我国有关部门大多是通过定期统计报表来收集统计资料的。新中国建立后的前三十年,抽样调查未得到足够的重视。全国范围内的应用主要是在 1955~1958 年以及 1962~1966 年两个短暂时期内,由国家统计局开展

的住户调查和农产量调查,其中50年代进行的农民家计调查覆盖了25个省、市、自治区的16468户,在60年代进行的类似调查覆盖了27个省、市、自治区的18000户。1963年进行的农产量调查覆盖了150个县15000个生产大队,实测地块4万个。在其后的两年内,调查的县、大队和实测地块分别增加了两倍与一倍。与此同时还组织了全国规模的城市职工家计调查。这些调查在“文化大革命”期间均告停顿。另一方面,建国头30年内,我国统计理论界中专门从事抽样理论与方法研究者更是寥若晨星。需要着重指出的是,我国统计界前辈许宝騄先生曾在60年前后在北京大学主持一个有关抽样调查的讨论班。根据许先生当时撰写的讲义整理出版的著作《抽样论》至今在我国统计界仍有很大影响。

党的十一届三中全会以后,我国实行了改革开放政策,社会经济面貌发生了根本的变化。在农村和部分城市企业中,逐渐实行了承包责任制,具有中国特色的社会主义市场经济逐步形成并取代过去单一的计划经济。以前可以通过报表制度获得的统计资料在新的条件下愈来愈困难。改革开放也导致人们观念的更新和思想活跃,各级领导和决策部门以及一些学术机构也需要了解掌握各阶层人们现状,及他们对社会中各种问题的看法和愿望,工商企业集团需要了解各消费阶层对其产品的需求与爱好。于是各种类型的抽样调查应运而生。因此进入80年代以来,我国抽样调查的应用与研究迎来了一个全面发展的新时期。

十多年来,国家统计局继续承担了国内最大量的抽样调查实际工作。为了适应新形势的需要,1984年国家统计局成立了城市社会经济调查总队与农村社会经济调查总队,在各省、市、自治区也建立了相应的队伍。其中“城调队”在146个市,80个县建队,编制4500人;“农调队”在857个县建队,编制8500人。两队分别在城市与农村进行定期的城市与农村住户的抽样调查,规模都数以万计。“农调队”还进行农村经济基本情况和农产量的抽样调查。1994年新成立了企业调查队。另外,国家人口普查办公室从1983年起,每年进行一次全国人口变动量的抽样调查。可以说,这些都是目前世界上进行的最大规模的抽样调查。尽管上述抽样调查的方法还比较单一,还存在某些具体问题有待解决,但发展势头是十分喜人的。最近国家统计局又提出今后的调查方法将以抽样调查为重点,以普查为框架,同时大力加强对抽样调查方法的应用研究。

除了国家统计局系统外,抽样调查在卫生部门与林业部门也有较长的应用历史。在卫生部门多次开展各种流行病学的抽样调查;在林业部

门则多将抽样调查用于动植物资源的估计上。在其他行业,例如交通部门,也从1988年起开展全国范围内定期的公路与水路交通运输量的抽样调查。此外,随着多种社会调查特别是市场调查的需要,各地先后成立了社会调查研究所,市场研究(调查)中心等半官方的或民间的以抽样调查为任务的机构。

十多年来,在国家有关部门组织下,还进行了多次不同目的的全国性的专项抽样调查,以下是其中影响较大的,不同领域的若干项目:

- 1) 全国高血压流行病学抽样调查(1979, 1991);
- 2) 全国结核病流行病学抽样调查(1979, 1985, 1990);
- 3) 全国千分之一妇女生育力调查(1982);
- 4) 全国粮食农药污染情况抽样调查(1984);
- 5) 中国成年人人体测量调查(1986);
- 6) 中国74个城镇人口迁移抽样调查(1986);
- 7) 全国科学研究与开发机构情况调查(1985, 1986);
- 8) 中国60岁以上老年人口抽样调查(1987);
- 9) 全国残疾人抽样调查(1987);
- 10) 中国儿童情况抽样调查(1987, 1992);
- 11) 为修订《服装号型》国家标准人体测量调查(1987);
- 12) 全国科技人员流动情况抽样调查(1987);
- 13) 全国专业技术人员情况抽样调查(1987);
- 14) 全国电视观众抽样调查(1987);
- 15) 全国1%人口抽样调查(1987);
- 16) 全国科技奖励工作抽样调查(1988);
- 17) 中央电视台收视率抽样调查(1989~);
- 18) 全国回国留学人员情况抽样调查(1989);
- 19) 石油系统干部情况抽样调查(1989);
- 20) 全国家用电器用户抽样调查(1990);
- 21) 公众对科学技术的态度抽样调查(1990, 1992);
- 22) 亚运会广播电视宣传效果抽样调查(1990);
- 23) 全国办公自动化抽样调查(1991);
- 24) 中国妇女社会地位抽样调查(1991);
- 25) 全国档案害虫分布与危害情况调查(1990~1991);

26) 中国家庭经济与生育调查(1991);

27) 全国服装消费行为调查(1992);

28) 国家卫生服务总调查(1993).

从以上列举的项目可以看到最近十多年来我国抽样调查应用的广泛和深入程度. 实际应用的需要也推动了对抽样调查方法与理论的研究.

上述项目中的大多数事先都进行了科学的抽样设计, 对多数项目也有与设计配套的数据处理方法. 这些都为进一步发展和推动抽样调查的应用与研究奠定了坚实的基础. 在本书的最后一章, 对上面所列的若干项目作为案例进行具体的介绍与分析.

第 2 章

简单随机抽样

简单随机抽样(simple random sampling)也称单纯随机抽样。从理论而言,这种抽样是最简单、最完善的,因此它构成抽样理论的基础。在实际中,简单随机抽样从样本抽取角度也是相当简单的。尽管就调查的实施而言,按照简单随机抽样可能存在许多实际困难,从而促使我们考虑其他抽样方法。例如分层抽样和多阶抽样,但即使在那些相对复杂一些的抽样中,层内抽样或最后一、两阶抽样也大量需要应用简单随机抽样。

在一些文献中,简单随机抽样又分为两种不同的情形:即不放回简单随机抽样(simple random sampling without replacement,简记为SRS WOR)及放回简单随机抽样(simple random sampling with replacement,简记为SRS WR)。在本书中,除非特别声明,我们将简单随机抽样都限制为前一种情形——不放回简单随机抽样。在本章前几节中只讨论这种情形。仅仅在§ 2.5中简要地讨论一下放回简单随机抽样,以便与一般的不放回情形进行比较。

§ 2.1 定义及实施方法

2.1.1 从一个有限总体中抽取所有可能的样本

设总体由 N 个抽样单元组成, N 已知, 欲在中间抽取包含 n 个抽样单元的样本, 称 n 为样本量(sample size), 是一个固定的数。为讨论样本的抽取方法, 我们从一个简单的实验例子出发, 研究从一个总体中可能取得的全部样本。

例 2.1 一个简单的实验例子: 考察一个 $N=8$ 的总体, 我们关心某个指标(变量) \mathcal{Y} , 第 i 个个体(单元)的指标值 Y_i 如表 2.1 所示。

从上述总体中抽取 $n=2$ 的样本, 可能样本的总数为

$$\binom{8}{2} = \frac{8!}{2! \times 6!} = 28.$$

表 2.1 $N=8$ 的一个人为总体

i	1	2	3	4	5	6	7	8
Y_i	4	6	8	10	7	3	8	5

表 2.2 从表 2.1 总体中抽取的 $n=2$ 全部可能的样本

样本号	样本单元	样本值	样本平均数	样本方差	样本号	样本单元	样本值	样本平均数	样本方差
1	Y_1, Y_2	4, 6	5.0	2.0	15	Y_3, Y_5	5, 7	6.0	2.0
2	Y_1, Y_3	4, 5	4.5	0.5	16	Y_3, Y_6	5, 3	4.0	2.0
3	Y_1, Y_4	4, 10	7.0	18.0	17	Y_3, Y_7	5, 8	6.5	4.5
4	Y_1, Y_5	4, 7	5.5	4.5	18	Y_3, Y_8	5, 5	5.0	0.0
5	Y_1, Y_6	4, 3	3.5	0.5	19	Y_4, Y_5	10, 7	8.5	4.5
6	Y_1, Y_7	4, 8	6.0	8.0	20	Y_4, Y_6	10, 3	6.5	24.5
7	Y_1, Y_8	4, 5	4.5	0.5	21	Y_4, Y_7	10, 8	9.0	2.0
8	Y_2, Y_3	6, 5	5.5	0.5	22	Y_4, Y_8	10, 5	7.5	12.5
9	Y_2, Y_4	6, 10	8.0	8.0	23	Y_5, Y_6	7, 3	5.0	8.0
10	Y_2, Y_5	6, 7	6.5	0.5	24	Y_5, Y_7	7, 8	7.5	0.5
11	Y_2, Y_6	6, 3	4.5	4.5	25	Y_5, Y_8	7, 5	6.0	2.0
12	Y_2, Y_7	6, 8	7.0	2.0	26	Y_6, Y_7	3, 8	5.5	12.5
13	Y_2, Y_8	6, 5	5.5	0.5	27	Y_6, Y_8	3, 5	4.0	2.0
14	Y_3, Y_4	5, 10	7.5	12.5	28	Y_7, Y_8	8, 5	6.5	4.5

每个样本包含的单元如表 2.2 所示. 为简便起见, 我们也用 Y_i 表示第 i 个单元.

注意: 表 2.2 中 28 个样本有一个共同的特点, 即同一单元在一个样本中都没有重复, 因此上述样本不包括 (Y_1, Y_1) 、 (Y_2, Y_2) 等. 如果样本中的两个单元是从总体中逐个抽取的, 这就意味着抽到第一个样本单元后, 不把它放回总体中, 而在其余 7 个单元中抽第二个样本单元, 这种抽样就是不放回抽样.

2.1.2 两个等价的定义

在实际问题中, 我们只需要按一定抽样方法从总体中抽取一个样本, 也即全部 $\binom{N}{n}$ 个可能样本中的一个. 一个简单且合理的原则是使这 $\binom{N}{n}$ 个样本每个被抽到的概率都相等. 满足这个条件的抽样方法就是 (不放回) 简单随机抽样.

定义 2.1 从总体中的 N 个单元中, 一次抽取 n 个单元, 使全部可能

的 $\binom{N}{n}$ 种不同的结果每种被抽到的概率都等于 $1/\binom{N}{n}$, 这种抽样称为简单随机抽样。按简单随机抽样, 抽到的样本称为简单随机样本。

当 N 、 n 比较大时, 按上述定义进行抽样是很不方便的。因为此时 $\binom{N}{n}$ 很大, 要列出全部可能的样本是不现实的。此时若按下面叙述的定义, 就较为容易实施了。

定义 2.2 从总体中的 N 个单元中, 逐个不放回地抽取单元, 每次抽取到尚未在样本(未入样)中的任何一个单元的概率都相等, 直到抽足 n 个单元为止, 这样所得的 n 个单元组成一个简单随机样本。

以上两个关于简单随机抽样的定义是等价的。为此我们只要证明每个包含有按定义 2.2 抽得的完全相同单元的样本被抽到的概率等于 $1/\binom{N}{n}$ 就可以了。

设按定义 2.2, 先后被抽中的单元号码为 i_1, i_2, \dots, i_n , 相应的样本值为 $Y_{i_1}, Y_{i_2}, \dots, Y_{i_n}$, 则抽到这样一个有序样本的概率为:

$$\begin{aligned} & P_r(Y_{i_1}, Y_{i_2}, \dots, Y_{i_n}) \\ &= P_r(Y_{i_1})P_r(Y_{i_2}|Y_{i_1})\cdots P_r(Y_{i_n}|Y_{i_1}, Y_{i_2}, \dots, Y_{i_{n-1}}) \\ &= \frac{1}{N} \cdot \frac{1}{N-1} \cdots \frac{1}{N-n+1} = \frac{(N-n)!}{N!}. \end{aligned}$$

实际上, 一个样本是不需要考虑其中单元抽取时的顺序的。一个包含有 n 个指定单元的样本, 其单元抽取的顺序共有 $n!$ 种不同的形式, 因此抽取到包含有这 n 个单元的样本的总概率为

$$\frac{(N-n)!n!}{N!} = \frac{1}{\binom{N}{n}}.$$

2.1.3 简单随机抽样的实施方法

根据定义, 简单随机抽样可用以下两种方法来实现:

方法 1(抽签法) 做 N 个签, 分别编上 $1 \sim N$ 号, 完全均匀混合后, 一次同时抽取 n 个签或一次抽取一个签但不把这个签放回, 接着抽第2个, 第3个, \dots , 直到抽足 n 个为止。上述两种程序实际上并无差别。所抽得的 n 个签上所示的号码即表示入样的单元号。

例如对例 2.1 的总体, 用抽签法抽一个 $n=2$ 的简单随机样本, 若抽中的签号为 3 与 2, 则 Y_3 与 Y_2 即为入样单元. 这相当于表 2.2 中的第 8 个样本.

方法 2 (随机数法) 利用随机数表, 随机数骰子或计算机产生的随机数进行抽样. 若利用计算机产生的随机数, 譬如说执行 BASIC 语言的 RAN 语句即可产生 $1 \sim N$ 范围(离散均匀分布)随机数(每个数每次出现的概率均为 $1/N$), 该数即表示入样的单元号. 若发生代表同一单元的随机数出现两次或两次以上, 则从第二次开始就弃去不用, 再抽下一个, 直到抽足 n 个不同的单元为止.

由于计算机产生的随机数实际上是伪随机数, 不是真正的随机数, 特别是直接采用一般现成程序时, 产生的随机数往往不能保证其随机性. 因此我们推荐使用随机数表或用随机数骰子产生的随机数, 特别是在样本量 n 比较大时. 利用随机数表或随机数骰子进行简单随机抽样的具体步骤将在 § 2.6 中详述.

§ 2.2 估计量及其性质

2.2.1 简单估计及其无偏性

我们用大写字母与小写字母分别表示有关总体与样本的量, 例如总体关于变量 \mathcal{Y} 的 N 个值记为 Y_1, Y_2, \dots, Y_N , 而

$$Y = \sum_{i=1}^N Y_i$$

和

$$\bar{Y} = \frac{Y}{N} = \frac{1}{N} \sum_{i=1}^N Y_i$$

分别表示总体总和及总体均值. 设 $Y_{i_1}, Y_{i_2}, \dots, Y_{i_n}$ 是从总体中抽取的一个样本量为 n 的简单随机样本, 其中 (i_1, i_2, \dots, i_n) 是 $(1, 2, \dots, N)$ 的一个子集. 根据前面的约定, 也为了简化足标起见, 将该样本重新记为 y_1, y_2, \dots, y_n , 于是

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (2.1)$$

即是样本平均数或称样本均值.

对简单随机抽样, 在没有对总体信息可以利用的情况下, 对 \bar{Y} 与 Y 的估计分别取为

$$\hat{\bar{Y}} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad (2.2)$$

$$\hat{Y} = N\bar{y} = \frac{N}{n} \sum_{i=1}^n y_i. \quad (2.3)$$

这种估计称为简单线性估计 (simple linear estimate), 简称为简单估计.

为讨论简单估计的性质, 首先证明以下两个引理:

引理 2.1 从大小为 N 的总体中抽取一个样本量为 n 的简单随机样本, 则总体中每个特定单元入样的概率为 n/N , 两个特定单元都入样的概率为 $\frac{n(n-1)}{N(N-1)}$.

证明 在全部可能的 $\binom{N}{n}$ 个样本中, 包含某个特定单元 Y_i 的样本数为 $\binom{N-1}{n-1}$ 个, 同时包含两个特定单元 Y_i, Y_j 的样本数为 $\binom{N-2}{n-2}$ 个, 而每个样本被抽到的概率都为 $1/\binom{N}{n}$, 因而每个单元入样的概率为

$$\binom{N-1}{n-1} / \binom{N}{n} = \frac{n}{N};$$

两个不同单元同时入样的概率为

$$\binom{N-2}{n-2} / \binom{N}{n} = \frac{n(n-1)}{N(N-1)}.$$

n/N 称为抽样比 (sampling fraction), 记为 f .

引理 2.2 从大小为 N 的总体中抽取一个样本量为 n 的简单随机样本, 对总体中的每个单元 Y_i , 引进随机变量 a_i 如下:

$$a_i = \begin{cases} 1, & \text{若 } Y_i \text{ 入样;} \\ 0, & \text{若 } Y_i \text{ 不入样} \end{cases} \quad (i=1, 2, \dots, N). \quad (2.4)$$

则

$$E(a_i) = \frac{n}{N} = f \quad (i=1, 2, \dots, N), \quad (2.5)$$

$$V(a_i) = \frac{n}{N} \cdot \frac{N-n}{N} = f(1-f) \quad (i=1, 2, \dots, N), \quad (2.6)$$

$$\text{Cov}(a_i, a_j) = -\frac{n}{N(N-1)} \left(1 - \frac{n}{N}\right) = -\frac{f(1-f)}{N-1} \quad (i, j=1, 2, \dots, N; i \neq j). \quad (2.7)$$

证明 显然, 每个 a_i 都服从二点分布, 根据引理 2.1, 有

$$E(a_i) = \frac{n}{N} = f, \quad E(a_i a_j) = \frac{n(n-1)}{N(N-1)} \quad (i \neq j).$$

因而

$$V(a_i) = f(1-f),$$

$$\text{Cov}(a_i, a_j) = E(a_i a_j) - E(a_i)E(a_j) = \frac{f(1-f)}{N-1}.$$

定理 2.1 对简单随机抽样, 作为 \bar{Y} 及 Y 的简单估计 \bar{y} 及 $\hat{Y} = N\bar{y}$ 都是无偏的, 即

$$E(\bar{y}) = \bar{Y}, \quad (2.8)$$

$$E(N\bar{y}) = Y. \quad (2.9)$$

我们只需证明其中一个结论, 例如(2.8)式即可. 下面我们给出三种证明, 每种证明的思想和方法都是具有启发意义的.

证明 1) 根据有限总体数学期望的含义, 有

$$E(\bar{y}) = \sum \frac{\bar{y}}{\binom{N}{n}},$$

这里求和是对全部可能的 $\binom{N}{n}$ 个不同的简单随机样本求的, \bar{y} 是每个样本的均值, 而每个样本被抽中的概率都为 $1/\binom{N}{n}$. 注意到对特定的总体单元 Y_i , 出现在不同样本中的次数为 $\binom{N-1}{n-1}$, 因此

$$\sum \bar{y} = \frac{1}{n} \sum (y_1 + y_2 + \cdots + y_n) = \frac{1}{n} \binom{N-1}{n-1} \sum_{i=1}^N Y_i,$$

$$\text{于是} \quad E(\bar{y}) = \frac{1}{n} \binom{N-1}{n-1} \sum_{i=1}^N Y_i / \binom{N}{n} = \frac{1}{N} \sum_{i=1}^N Y_i = \bar{Y},$$

证明 2) 由于总体中每小单元 Y_i 出现在全部可能的简单随机样本的次数都相等, 因此 $E\left[\sum_{i=1}^n y_i\right]$ 作为对所有可能样本求平均, 它必定是 $\sum_{i=1}^N Y_i$ 的倍数. 根据求和中的单元个数计算, 这个倍数恰为 $\frac{n}{N}$, 因而

$$E(y) = \frac{1}{n} \frac{n}{N} \sum_{i=1}^N Y_i = \frac{1}{N} \sum_{i=1}^N Y_i = \bar{Y}.$$

上述证明方法称为“对称性论证”(argument of symmetry), 这种方法对证明简单随机抽样的有关性质是十分方便的.

证明 3) (Cornfield) 引进随机变量

$$a_i = \begin{cases} 1, & \text{若 } Y_i \text{ 入样} \\ 0, & \text{否则} \end{cases} \quad (i=1, 2, \dots, N),$$

则 \bar{y} 可表达为

$$\bar{y} = \frac{1}{n} \sum_{i=1}^N a_i Y_i,$$

其中 $Y_i (i=1, 2, \dots, N)$ 都是常数, 故

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^N Y_i E(a_i) = \frac{1}{n} \frac{n}{N} \sum_{i=1}^N Y_i = \bar{Y}. \blacksquare$$

对于例 2.1, 表 2.2 中给出了全部 28 个可能的简单随机样本的平均数 \bar{y} , 读者不难验明这 28 个平均数的平均数等于总体均值 $\left(\frac{12}{7}\right)$, 说明 \bar{y} 是无偏的.

2.2.2 估计量的方差与协方差

一、 \bar{y} 的方差 $V(\bar{y})$

为表达 \bar{y} 的方差, 我们先定义总体的方差. 按一般定义, 有限总体的方差为

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2, \quad (2.10)$$

但为了在大多数情形使公式的表达更为简练, 在本书中我们用

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad (2.11)$$

来表示总体方差. 这种表示方式在用方差分析法处理时尤为方便.

定理 2.2 对简单随机抽样, \bar{y} 的方差为

$$V(\bar{y}) = \frac{S^2}{n} (1-f). \quad (2.12)$$

证明 1) (对称性论证法)

$$\begin{aligned} V(\bar{y}) &= E(\bar{y} - \bar{Y})^2 = \frac{1}{n^2} E[n(\bar{y} - \bar{Y})]^2 \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^n (y_i - \bar{Y})\right]^2 \\ &= \frac{1}{n^2} \left\{ E\left[\sum_{i=1}^n (y_i - \bar{Y})^2\right] + 2E\left[\sum_{i < j}^n (y_i - \bar{Y})(y_j - \bar{Y})\right] \right\}. \end{aligned}$$

根据对称性论证法, 我们有

$$E\left[\sum_{i=1}^n (y_i - \bar{Y})^2\right] = \frac{n}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

$$E \left[\sum_{i < j}^n (y_i - \bar{Y})(y_j - \bar{Y}) \right] = \frac{n(n-1)}{N(N-1)} \sum_{i < j}^N (Y_i - \bar{Y})(Y_j - \bar{Y}),$$

故

$$\begin{aligned} V(\bar{y}) &= \frac{1}{nN} \left[\sum_{i=1}^N (Y_i - \bar{Y})^2 + 2 \frac{n-1}{N-1} \sum_{i < j}^N (Y_i - \bar{Y})(Y_j - \bar{Y}) \right] \\ &= \frac{1}{nN} \left\{ \left(1 - \frac{n-1}{N-1} \right) \sum_{i=1}^N (Y_i - \bar{Y})^2 + \frac{n-1}{N-1} \left[\sum_{i=1}^N (Y_i - \bar{Y}) \right]^2 \right\} \\ &= \frac{1}{n} \cdot \frac{N-n}{N} \cdot \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{S^2}{n} (1-f). \end{aligned}$$

证明 2) (Cornfield 法) 仍引进随机变量

$$a_i = \begin{cases} 1, & \text{若 } Y_i \text{ 入样} \\ 0, & \text{否则} \end{cases} \quad (i = 1, 2, \dots, N).$$

于是根据引理 2.2, 有

$$\begin{aligned} V(\bar{y}) &= V \left[\frac{1}{n} \sum_{i=1}^N a_i Y_i \right] \\ &= \frac{1}{n^2} \left[\sum_{i=1}^N Y_i^2 V(a_i) + 2 \sum_{i < j}^N Y_i Y_j \text{Cov}(a_i, a_j) \right] \\ &= \frac{1-f}{nN} \left[\frac{N}{N-1} \sum_{i=1}^N Y_i^2 - \frac{1}{N-1} \left(\sum_{i=1}^N Y_i \right)^2 \right] \\ &= \frac{1-f}{n(N-1)} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{S^2}{n} (1-f). \quad \blacksquare \end{aligned}$$

推论 对于简单随机抽样, $\hat{Y} = N\bar{y}$ 的方差为

$$V(\hat{Y}) = \frac{S^2}{n} N(N-n) = \frac{N^2 S^2}{n} (1-f). \quad (2.13)$$

定理 2.2 中的

$$1-f = \frac{N-n}{N} \quad (2.14)$$

称为有限总体校正系数 (finite population correction, 简记为 FPC). 这是因为对无限总体中的抽样, $V(\bar{y})$ 应等于 $\sigma^2/n \approx S^2/n$ (参见 § 2.5 中对放回简单随机抽样的讨论), 因此从有限总体中抽得的简单随机样本均值的方差要比从无限总体中独立样本均值的方差小, 两者相差 $1-f$ 这样一个因子. 当抽样比 f 很小时 (例如 $f < 0.05$), 因子 $1-f$ 可以忽略不计. 定理 2.2 告诉我们, 影响 \bar{y} 精度的主要是样本量 n 的大小, 而不是抽样比 f . 这一点对初学者尤其要引起注意.

二、两个估计量 \bar{y} 、 \bar{x} 的协方差

若总体中的每个单元都有两个指标 Y_i 与 X_i , 记 \bar{y} 、 \bar{x} 为相应的样本

均值, \bar{Y} 与 \bar{X} 分别为总体均值, 则可定义 \bar{y} 与 \bar{x} 的协方差如下:

$$\text{Cov}(\bar{y}, \bar{x}) = E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}). \quad (2.15)$$

定理 2.3 对简单随机抽样, 有

$$\text{Cov}(\bar{y}, \bar{x}) = \frac{1-f}{n} S_{yx} \quad (2.16)$$

其中

$$S_{yx} = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}) \quad (2.17)$$

是总体协方差.

证明 1) 用对称论证法

$$\begin{aligned} \text{Cov}(\bar{y}, \bar{x}) &= E[(\bar{y} - \bar{Y})(\bar{x} - \bar{X})] \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^n (y_i - \bar{Y})\right]\left[\sum_{j=1}^n (x_j - \bar{X})\right] \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^n (y_i - Y)(x_i - \bar{X}) + \sum_{i=1}^n \sum_{j \neq i}^n (y_i - Y)(x_j - \bar{X})\right] \\ &= \frac{1}{n^2} \left[\frac{n}{N} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}) \right. \\ &\quad \left. + \frac{n(n-1)}{N(N-1)} \sum_{i=1}^N \sum_{j \neq i}^N (y_i - \bar{Y})(x_j - \bar{X}) \right] \\ &= \left[\frac{1}{nN} - \frac{n}{N(N-1)} \frac{1}{n} \right] \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}) \\ &\quad + \frac{n-1}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N (y_i - \bar{Y})(x_j - \bar{X}) \\ &= \frac{N-n}{nN(N-1)} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}) \\ &\quad + \frac{n-1}{N(N-1)} \left[\sum_{i=1}^N (y_i - \bar{Y}) \right] \left[\sum_{j=1}^N (x_j - \bar{X}) \right] \\ &= \frac{1-f}{n} S_{yx}. \end{aligned}$$

证明 2) 令 $u_i = y_i + x_i$, 记 \bar{u}, \bar{U} 分别为样本均值与总体均值, 则 $\bar{u} = \bar{y} + \bar{x}$, $\bar{U} = \bar{Y} + \bar{X}$.

$$\begin{aligned} V[(\bar{y} - \bar{Y}) + (\bar{x} - \bar{X})] &= V(\bar{y} - \bar{Y}) + V(\bar{x} - \bar{X}) \\ &\quad + 2 \text{Cov}[(\bar{y} - \bar{Y}), (\bar{x} - \bar{X})] \\ &= V(\bar{y}) + V(\bar{x}) + 2 \text{Cov}(\bar{y}, \bar{x}). \end{aligned}$$

$$\text{Cov}(\bar{y}, \bar{x}) = \frac{1}{2} [V(\bar{u}) - V(\bar{y}) - V(\bar{x})]$$

$$\begin{aligned}
&= \frac{1}{2} \cdot \frac{1-f}{n} (S_u^2 - S_y^2 - S_x^2) \\
&= \frac{1}{2} \cdot \frac{1-f}{n} \cdot \frac{1}{N-1} \left[\sum_{i=1}^N (y_i - \bar{y} - \bar{Y} - \bar{X})^2 \right. \\
&\quad \left. - \sum_{i=1}^N (y_i - \bar{Y})^2 - \sum_{i=1}^N (x_i - \bar{X})^2 \right] \\
&= \frac{1}{2} \cdot \frac{1-f}{n} \cdot \frac{2}{N-1} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X}) \\
&= \frac{1-f}{n} S_{yx}. \blacksquare
\end{aligned}$$

2.2.3 方差与协方差的估计

在实际问题中, 总体的方差与协方差都是未知的, 因此为了得到估计量方差或协方差的估计, 必须对总体的方差与协方差进行估计.

定理 2.4 简单随机样本的方差

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2.18)$$

是总体方差 S^2 的无偏估计.

证明 s^2 可改写成

$$\begin{aligned}
s^2 &= \frac{1}{n-1} \sum_{i=1}^n [(y_i - \bar{Y}) - (\bar{y} - \bar{Y})]^2 \\
&= \frac{1}{n-1} \left[\sum_{i=1}^n (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \right].
\end{aligned}$$

根据对称性论证及定理 2.2, 有

$$\begin{aligned}
E \left[\sum_{i=1}^n (y_i - \bar{Y})^2 \right] &= \frac{n}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{n(N-1)}{N} S^2, \\
E [n(\bar{y} - \bar{Y})^2] &= nV(\bar{y}) = \frac{N-n}{N} S^2.
\end{aligned}$$

所以
$$E(s^2) = \frac{S^2}{(n-1)N} [n(N-1) - (N-n)] = S^2. \blacksquare$$

作为例子, 对例 2.1, 总体方差 $S^2 = \frac{96}{7} = 5.1429$, 而表 2.2 中全部可能样本方差的平均数 $E(s^2) = 144/28 = 5.1429$. 两者相等.

推论 对于简单随机抽样

$$v(\bar{y}) \triangleq s_y^2 = \frac{S^2}{n} (1-f) \quad (2.19)$$

与

$$v(\hat{Y}) = s_Y^2 \frac{N^2 s^2}{n} (1-f) \quad (2.20)$$

分别是 $V(\bar{y})$ 与 $V(\hat{Y})$ 的无偏估计.

在获得估计量方差估计后, 即可根据 1.2.5 段中的方法来构造总体参数的(近似)置信区间. 例如对总体平均数 \bar{Y} , 一个置信水平为 $1 - \alpha$ 的近似置信区间为

$$\left[\bar{y} - u_\alpha \sqrt{\frac{1-f}{n}} s, \bar{y} + u_\alpha \sqrt{\frac{1-f}{n}} s \right]. \quad (2.21)$$

例 2.2 某市区共有 4328 户. 为调查该区居民的收入情况, 用简单随机抽样方法从中抽取 30 户, 登记了每户的月收入 y_i , 具体数据如表 2.3 所示. 试估计该区居民的平均月收入 \bar{Y} , 并求它的置信水平为 95% 的近似置信区间.

表 2.3 30 户居民的月收入调查

序号 i	月收入 y_i (元)	序号 i	月收入 y_i (元)
1	670	16	716
2	760	17	456
3	510	18	904
4	676	19	928
5	764	20	664
6	494	21	930
7	724	22	760
8	840	23	734
9	580	24	604
10	574	25	554
11	768	26	656
12	690	27	684
13	880	28	760
14	560	29	496
15	650	30	920

这里 $N = 4328$, $n = 30$.

根据表 2.3, 计算得

$$\begin{aligned} \bar{y} &= \frac{1}{30} \sum_{i=1}^{30} y_i = \frac{1}{30} \times 20886 = 696.20, \\ s^2 &= \frac{1}{30-1} \sum_{i=1}^{30} (y_i - \bar{y})^2 = \frac{1}{29} \left[\sum_{i=1}^{30} y_i^2 - n\bar{y}^2 \right] \\ &= \frac{1}{29} \times 536994.8 = 18517.06. \end{aligned}$$

$$v(\bar{y}) = \frac{1}{30} \times \left[1 - \frac{30}{4928} \right] \times 18517.06 = 612.96,$$

$$s(\bar{y}) = \sqrt{v(\bar{y})} = 24.76.$$

因而该区居民户平均月收入 Y 的估计为 696.20 元, 而它的 95% 的近似置信区间为:

$$696.20 \pm 1.96 \times 24.76$$

即(647.67 元, 744.73 元).

与定理 2.4 类似, 在有两个指标的情形, 我们有以下定理:

定理 2.5 简单随机样本的协方差

$$s_{yx} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \quad (2.22)$$

是总体协方差 S_{yx} 的无偏估计.

证明留给读者作练习. 根据定理 2.5, 可以构造 $\text{Cov}(\bar{y}, \bar{x})$ 的无偏估计.

2.2.4 简单估计的优良性及可以进一步改进的途径

简单(线性)估计不仅有简洁的形式, 而且也具有某些优良性质.

Neyman 与 David (1938) 证明了若将 y_1, y_2, \dots, y_n 视为 n 个随机变量, 具有公共均值 \bar{Y} 及相同方差与协方差, 则简单估计 \bar{y} 是一致最小方差的线性无偏估计. 这是 Markoff 定理的一种特殊情形.

Horvitz 与 Thompson (1952) 证明了在形如 $\hat{Y} = \sum_{i=1}^n w_i y_i$ 的线性估计类里, 若要求对所有的 w_i 都相等, 那么 $w_i = \frac{n}{N}$ 是 \hat{Y} 为无偏的充分必要条件. 这表明, 此时 \bar{y} 是唯一满足条件的无偏估计. 如果权 w_i 只取决于入样的顺序 d , 记第 d 次入样的样本单元为 $y_{(d)}$, 则 \bar{y} 在形如 $\sum_{d=1}^n w_d y_{(d)}$ 的线性估计类中方差最小.

若 y_i 的权不仅依赖于 i 而且也与样本中的其他单元有关, 记为 w_{is} , 则 Godambe (1955) 证明了对于所有总体在形如 $\sum_{i=1}^n w_{is} y_i$ 的估计类中不存在最小方差的无偏估计. 对于某些总体, 选取一定的 w_{is} , 就有可能构造出比简单估计方差更小的估计量.

事实上, 如果我们对总体的特性有一定了解, 即使在简单随机抽样范围内, 也常可找到优于简单估计的其他估计形式, 看下面的例子:

例 2.3 若在总体中, 已知某个单元, 设为 Y_1 , 很小; 而另一个单元, 设为 Y_N , 很大, 则 Särndal 提出对 \bar{Y} 的如下估计量:

$$\hat{\bar{Y}}_s = \begin{cases} y + c, & \text{若样本中包含 } Y_1 \text{ 而不包含 } Y_N; \\ y - c, & \text{若样本中包含 } Y_N \text{ 而不包含 } Y_1; \\ \bar{y}, & \text{其他情形.} \end{cases} \quad (2.23)$$

其中 c 是常数, 则 $\hat{\bar{Y}}_s$ 是无偏的, 且

$$V(\hat{\bar{Y}}_s) = (1-f) \left[\frac{S^2}{n} - \frac{2c}{N-1} (Y_N - Y_1 - nc) \right], \quad (2.24)$$

因而当 c 满足

$$0 < c < \frac{1}{n} (Y_N - Y_1)$$

时, $V(\hat{\bar{Y}}_s) < V(\bar{y})$.

为证明 $\hat{\bar{Y}}_s$ 的无偏性, 我们引进随机变量 a_i :

$$a_i = \begin{cases} 1, & \text{若 } Y_i \text{ 入样} \\ 0, & \text{否则} \end{cases} \quad (i=1, 2, \dots, N).$$

则 $\hat{\bar{Y}}_s$ 可表成

$$\hat{\bar{Y}}_s = \frac{1}{n} \left[a_1(Y_1 + nc) + a_N(Y_N - nc) + \sum_{i=2}^{N-1} a_i Y_i \right]. \quad (2.25)$$

由于 $E(a_i) = \frac{n}{N}$ ($i=1, 2, \dots, N$), 故

$$\begin{aligned} E(\hat{\bar{Y}}_s) &= \frac{1}{n} \cdot \frac{n}{N} \left(Y_1 + nc + Y_N - nc + \sum_{i=2}^{N-1} Y_i \right) \\ &= \frac{1}{N} \sum_{i=1}^N Y_i = \bar{Y}. \end{aligned}$$

为证明(2.24)式, 我们先给出总体方差 S^2 的一种表达式:

$$\begin{aligned} S^2 &= \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 \\ &= \frac{1}{N-1} \left[\sum_{i=1}^N Y_i^2 - \frac{1}{N} \left(\sum_{i=1}^N Y_i \right)^2 \right] \\ &= \frac{1}{N-1} \left[\sum_{i=1}^N Y_i^2 - \frac{1}{N} \left(\sum_{i=1}^N Y_i^2 + 2 \sum_{i < j}^N Y_i Y_j \right) \right] \\ &= \frac{1}{N} \sum_{i=1}^N Y_i^2 - \frac{2}{N(N-1)} \sum_{i < j}^N Y_i Y_j. \end{aligned} \quad (2.26)$$

对(2.25)式求方差, 利用 a_i 的方差与协方差的表达式((2.6)与(2.7))

式), 有

$$\begin{aligned}
 V(\hat{\bar{Y}}_c) &= \frac{1}{n^2} \left\{ \frac{n}{N} (1-f) \left[(Y_1 + nc)^2 + (Y_N - nc)^2 + \sum_{i=2}^{N-1} Y_i^2 \right] \right. \\
 &\quad - \frac{2n}{N(N-1)} (1-f) \left[(Y_1 + nc)(Y_N - nc) + \sum_{i=2}^{N-1} Y_1 Y_i \right. \\
 &\quad \left. \left. + nc \sum_{i=2}^{N-1} Y_i - nc \sum_{i=2}^{N-1} Y_i + \sum_{i=2, j=2, \dots, N}^N Y_i Y_j \right] \right\} \\
 &= (1-f) \left[\frac{1}{nN} \sum_{i=1}^N Y_i^2 - \frac{2}{nN(N-1)} \sum_{i < j}^N Y_i Y_j \right] \\
 &\quad - (1-f) \left[\frac{2c}{N} (Y_N - Y_1 - nc) \right. \\
 &\quad \left. + \frac{2c}{N(N-1)} (Y_N - Y_1 - nc) \right] \\
 &= (1-f) \left[\frac{S^2}{n} - \frac{2c}{N-1} (Y_N - Y_1 - nc) \right].
 \end{aligned}$$

下面我们用一个简单实验例子来进一步说明问题. 设 $N=8$ 的一个总体, 其单元数值为:

$$1, 4, 5, 5, 6, 6, 8, 13.$$

从中抽取 $n=4$ 的简单随机样本, 则不难验明 $V(\bar{y})=1.5$, 而当 $0 < c < 8$ 时就有 $V(\hat{\bar{Y}}_c) < V(\bar{y})$. 例如当 $c=1$ 或 2 时, $V(\hat{\bar{Y}}_c)=0.357$, 当 $c=1.5$ 时, $V(\hat{\bar{Y}}_c)$ 达到最小值 0.214 .

对这个特殊例子还可以考虑另一种不同于简单随机抽样的抽样方法: 每一个样本均包含 $Y_1=1$ 与 $Y_8=13$, 同时在另外 6 个单元中按简单随机抽样抽取一个 $n'=2$ 的样本, 记这个样本的平均数为 \bar{y}_2 , 考虑估计量

$$\bar{y}_{st} = \frac{1}{8} [Y_1 + Y_8 + 6\bar{y}_2],$$

则 \bar{y}_{st} 也是 \bar{Y} 的无偏估计量, 而且 $V(\bar{y}_{st})=0.350$ 也小于 $V(\bar{y})$.

这个例子说明了为提高简单随机抽样简单估计精度的两种途径. 一是改变抽样方法, 上面提到的 \bar{y}_{st} 即是一种特殊分层抽样的简单估计. 在下一章中将详细讨论分层抽样. 第二种途径是对简单随机样本利用总体的一定信息采用有别于简单估计的另外估计方法, 正如 $\hat{\bar{Y}}_c$. 当总体中每个单元还有辅助变量可以利用时, 还可采用精度更高的比估计、回归估计等. 这将在第 4 章中详细讨论.

§ 2.3 总体比例的估计与对子总体的估计

2.3.1 总体比例(百分率)的估计

设总体中的 N 个单元按某种特征分成两类, 一类具有这种特征, 另一类不具有这种特征. 我们的目的是估计总体中具有这种特征的单元在全体单元中所占的比例 P 或总体中具有这种特征的单元总数 A . 例如男性的比例、患结核病人数的比例、选民在一次选举中的投票率等等.

若对每个单元, 定义指标值

$$Y_i = \begin{cases} 1, & \text{若第 } i \text{ 个单元具有所考虑的特征;} \\ & (i = 1, 2, \dots, N) \\ 0, & \text{否则.} \end{cases} \quad (2.27)$$

则有

$$A = \sum_{i=1}^N Y_i = Y, \quad P = \frac{A}{N} = Y. \quad (2.28)$$

因而对总体比例的估计即可化成上节讨论的一般情形即总体均值的估计.

定理 2.6 若 a 是样本量为 n 的简单随机样本中具有所考虑特征的单元数, 则样本比例

$$p = \frac{a}{n} \quad (2.29)$$

是总体比例 $P = \frac{A}{N}$ 的无偏估计, 且

$$V(p) = \frac{PQ}{n} \cdot \frac{N-n}{N-1}, \quad (2.30)$$

其中

$$Q \triangleq 1 - P = \frac{N-A}{N}.$$

证明 引进 Y_i 如前, 则 $p = \bar{y}$, 根据定理 2.1, 即有 $E(p) = P$. 另一方面, 此时总体方差为:

$$\begin{aligned} S^2 &= \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i^2 - N\bar{Y}^2) \\ &= \frac{1}{N-1} (NP - NP^2) = \frac{N}{N-1} P(1-P) \\ &= \frac{N}{N-1} PQ. \end{aligned} \quad (2.31)$$

将上式代入定理 2.2 中的(2.12)式即得(2.30)式. ■

推论 $\hat{A} = Np$ 是 A 的无偏估计, 且

$$V(\hat{A}) = \frac{N^2 PQ}{n} \cdot \frac{N-n}{N-1}. \quad (2.32)$$

定理 2.7 对简单随机抽样

$$v(p) \triangleq S_p^2 = \frac{N-n}{N(n-1)} pq \cdot \frac{1-f}{n-1} pq \quad (2.33)$$

是 $V(p)$ 的一个无偏估计, 其中 $q = 1-p$.

证明 由(2.27)式不难验证, 此时样本方差为

$$s^2 = \frac{n}{n-1} pq. \quad (2.34)$$

从而由定理 2.4 的推论即得证. ■

(2.33) 式表明: 当 $1-f \approx 1$ 时, $\frac{pq}{n-1}$ 是 $V(p)$ 的无偏估计, 而 $\frac{pq}{n}$ 则是有偏的.

推论 $V(\hat{A})$ 的一个无偏估计是

$$v(\hat{A}) = \frac{N(N-n)}{n-1} pq. \quad (2.35)$$

有了 $v(p)$ 或 $v(\hat{A})$, 即可构造 P 或 A 的置信区间, 当 n 很大时, 可以用前述通用的近似方法. 也即 P 的置信度为 $1-\alpha$ 的近似置信区间为

$$\left[p - u_\alpha \sqrt{\frac{(1-f)pq}{n-1}}, p + u_\alpha \sqrt{\frac{(1-f)pq}{n-1}} \right]. \quad (2.36)$$

由于 $Y_i(y_i)$ 的取值仅是 0 和 1 两个值, 因此 $\alpha(p)$ 的实际分布 (当 N 很大时) 为二项分布, 是离散的. 当 n 不是很大时, 应考虑作连续性修正. 此时 P 的置信区间可修正为:

$$\left[p - \left(u_\alpha \sqrt{\frac{(1-f)pq}{n-1}} + \frac{1}{2n} \right), p + \left(u_\alpha \sqrt{\frac{(1-f)pq}{n-1}} + \frac{1}{2n} \right) \right]. \quad (2.37)$$

而 A 的置信限可用同样情况下 P 的置信限乘以 N 而得到.

例 2.4 某大学有学生 5620 人. 为了解现任学生会主席在换届选举中连任的可能性, 在学生中用简单随机抽样调查了 300 名学生, 其中有 187 人支持主席连任. 试估计该校学生支持主席连任的比例及总人数.

解 这里 $N = 5620$, $n = 300$, $1-f = 0.9466$,

$$\alpha = 187, \quad p = \frac{\alpha}{n} = 0.6233, \quad q = 0.3767,$$

$$\hat{A} = Np = 3502.9 \approx 3503,$$

$$v(p) \triangleq s_p^2 = \frac{1-f}{n-1} pq = 7.4334 \times 10^{-4},$$

$$s_p = 0.02726,$$

$$s_{N_p} = 153.2 \approx 153.$$

于是 P 与 A 的 90% 的置信限及置信区间 ($u_{.10} = 1.64$) 分别为:

$P: 0.6233 \pm 1.64 \times 0.02726$ 即 $(0.5786, 0.6680)$;

$A: 3503 \pm 1.64 \times 153$ 即 $(3252, 3754)$.

若考虑连续性修正, 即用 (2.37) 式, 相应的置信区间为:

$P: (.5769, 0.6697); \quad A: (3242, 3764).$

2.3.2 子总体的估计

有时总体单元可以按一种或几种可辨别的特征划分成若干个子总体 (subpopulations). 例如在调查对象为人时, 按性别或年龄段划分; 在调查对象为企业时, 按规模大小或所有制性质划分. 我们关心的是对这些子总体参数的估计. 在有些文献上, 例如联合国统计分委员会将这种感兴趣的子总体称为研究域 (domains of study). 在对子总体 (研究域) 进行估计时, 每个子总体的大小不一定是已知的. 因此对单元的划分通常只能在样本中进行.

令 N_j 是第 j 个子总体的大小 (设它未知). 从总体中抽取一个样本量为 n 的简单随机样本, 样本中属于第 j 个子总体的单元数为 n_j . 则这 n_j 个单元可看成是从大小为 N_j 的 (子) 总体中抽取的一个简单随机样本. 与一般情形不同的是, 这里的 n_j 并不能事先确定.

对第 j 个子总体, 记它的第 i 个单元的指标值为 $Y_i^{(j)}$, 样本中的指标值为 $y_i^{(j)}$, 根据定理 2.1, $\bar{y}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_i^{(j)}$ 是子总体均值 $\bar{Y}^{(j)} = \sum_{i=1}^{N_j} Y_i^{(j)} / N_j$ 的无偏估计. 而它的方差为

$$(1-f_j) \frac{S_j^2}{n_j} = \left(1 - \frac{n_j}{N_j}\right) \frac{S_j^2}{n_j},$$

其中子总体方差 S_j^2 可用样本方差 $\frac{1}{n_j-1} \sum_{i=1}^{n_j} (y_i^{(j)} - \bar{y}^{(j)})^2$ 估计. 为估计 n_j, N_j , 注意到若将属于第 j 个子总体看作是总体单元的一个特征, 则比例 N_j/N 可用相应的样本比例 n_j/n 估计, 即

$$E\left(\frac{n_j}{n}\right) = \frac{N_j}{N}. \quad (2.38)$$

N_j 虽未知, 但是一个常数, 故上式可改写为

$$E\left(\frac{n_j}{N_j}\right) = \frac{n}{N}. \quad (2.39)$$

因而可用 $f = \frac{n}{N}$ 来估计 $f_j = \frac{n_j}{N_j}$. 于是 $V(\bar{y}^{(j)})$ 可用下式估计:

$$v(\bar{y}^{(j)}) = \frac{1}{n_j} \cdot \frac{f}{n_j - 1} \sum_{i=1}^{n_j} (y_i^{(j)} - \bar{y}^{(j)})^2. \quad (2.40)$$

下面我们讨论另一种常见的关于子总体的估计问题, 即估计子总体某个指标 \mathscr{Z} 的总和. 这个问题可以化为 2.3.1 段中的问题进行处理. 此时所考虑的子总体的大小 N_j 即是 2.3.1 段中的 A .

对总体中的每个单元, 定义指标值 \mathscr{Z} 如下:

$$Y_i = \begin{cases} Z_i, & \text{若第 } i \text{ 个单元属于所考虑的子总体;} \\ 0, & \text{否则.} \end{cases} \quad (2.41)$$

于是

$$Y = \sum_{i=1}^N Y_i = \sum_{i=1}^A Z_i = Z$$

即是我们需要估计的参数.

设对总体抽取的简单随机样本中属于该子总体的单元数为 a (即 n_j), 记

$$P = \frac{A}{N}, \quad Q = 1 - P, \quad (2.42)$$

$$p = \frac{a}{n}, \quad q = 1 - p. \quad (2.43)$$

则

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i = \frac{1}{N} \sum_{i=1}^A Z_i = \frac{A}{N} \cdot \frac{1}{A} \sum_{i=1}^A Z_i = PZ, \quad (2.44)$$

$$\begin{aligned} S_Y^2 &= \frac{1}{N-1} \left(\sum_{i=1}^N Y_i^2 - N\bar{Y}^2 \right) = \frac{1}{N-1} \left(\sum_{i=1}^A Z_i^2 - NP^2\bar{Z}^2 \right) \\ &= \frac{1}{N-1} \left[\sum_{i=1}^A (Z_i - Z)^2 + A\bar{Z}^2 - NP^2\bar{Z}^2 \right] \\ &= \frac{1}{N-1} \left[(A-1)S_Z^2 + N\bar{Z}^2 \left(\frac{A}{N} - P^2 \right) \right] \\ &= \frac{A-1}{N-1} S_Z^2 + \frac{N}{N-1} PQ\bar{Z}^2. \end{aligned} \quad (2.45)$$

根据上述的关系及记号, 即有以下定理:

定理 2.8 对简单随机抽样

$$\hat{Z} = \frac{N}{n} \sum_{i=1}^a z_i \quad (2.46)$$

是子总体总和 Z 的无偏估计, 其方差为

$$V(\hat{Z}) = \frac{N^2}{n}(1-f) \left[\frac{A-1}{N-1} S_z^2 + \frac{N}{N-1} PQ\bar{Z}^2 \right] \quad (2.47)$$

$$\approx \frac{N^2}{n}(1-f)(PS_z^2 + PQ\bar{Z}^2), \quad (2.47')$$

且

$$v(\hat{Z}) = \frac{N^2(1-f)}{n(n-1)} \left[\sum_{i=1}^a (z_i - \bar{z})^2 + npq\bar{z}^2 \right] \quad (2.48)$$

是 $V(\hat{Z})$ 的无偏估计. 这里 \bar{z} 是属于子总体的样本 z_i ($i=1, 2, \dots, a$) 的均值.

例 2.5 在一个有 23482 户的镇中抽一个 565 户的简单随机样本, 对每一住户调查住房的居住面积与住房性质(公房或私房), 基本结果如表 2.4 所示.

表 2.4 某镇的住户调查(面积单位: m^2)

住房性质	样本中的户数 $n^{(j)}$	平均户居住面积 $\bar{z}^{(j)}$	户居住面积标准差 $s_z^{(j)}$
公房	386	31.20	10.39
私房	179	24.52	5.62
合计	565		

分别估计该镇居民住公房、私房的比例 $P^{(j)}$, 户数 $A^{(j)}$ 及该镇公、私房的总居住面积 $Z^{(j)}$ 以及 $\hat{Z}^{(j)}$ 的标准差.

解 对子总体(1): 公房住户的估计:

$$p^{(1)} = \frac{n_1}{n} = \frac{386}{565} = 0.6832, \quad q^{(1)} = 1 - p^{(1)} = 0.3168,$$

$$\hat{A}^{(1)} = Np^{(1)} = 23482 \times 0.6832 = 16043,$$

$$\begin{aligned} \hat{Z}^{(1)} &= \frac{N}{n} \sum_{i=1}^{n_1} z_i^{(1)} = \frac{N}{n} n_1 \bar{z}^{(1)} = \frac{23482}{565} (31.20 \times 386) \\ &= 500528 (\text{m}^2), \end{aligned}$$

$$\begin{aligned} v(\hat{Z}^{(1)}) &= \frac{N}{n(n-1)} \left[\sum_{i=1}^{n_1} (z_i^{(1)} - \bar{z}^{(1)})^2 + np^{(1)}q^{(1)}\bar{z}^{(1)2} \right] \\ &= \frac{N}{n} [s_z^{(1)2}]^2 + \frac{N}{n-1} p^{(1)}q^{(1)}[\bar{z}^{(1)}]^2 \\ &= 23482 \times (0.209898 + 0.373562) = 13700.82, \end{aligned}$$

$$s(\hat{Z}^{(1)}) = \sqrt{v(\hat{Z}^{(1)})} = 117.05 (\text{m}^2).$$

对子总体(2): 私房住户估计:

$$p^{(2)} = \frac{n_2}{n} = \frac{179}{565} = 0.3168, \quad q^{(2)} = 1 - p^{(2)} = 0.6832,$$

$$\hat{A}^{(2)} = N p^{(2)} = 23482 \times 0.3168 = 7439,$$

$$\begin{aligned} \hat{Z}^{(2)} &= \frac{N}{n} \sum_{i=1}^{n_1} z_i^{(2)} = \frac{N}{n} n_2 \bar{z}^{(2)} = \frac{23482}{565} \times 24.52 \times 179 \\ &= 182412 (\text{m}^2), \end{aligned}$$

$$\begin{aligned} v(\hat{Z}^{(2)}) &= \frac{N}{n(n-1)} \left[\sum_{i=1}^{n_1} (z_i^{(2)} - \bar{z}^{(2)})^2 + n p^{(2)} q^{(2)} (\bar{z}^{(2)})^2 \right] \\ &= \frac{N}{n} [s_z^{(2)}]^2 + \frac{N}{n-1} p^{(2)} q^{(2)} [\bar{z}^{(2)}]^2 \\ &= 23482 \times (0.055902 + 0.230725) = 6730.57, \end{aligned}$$

$$s(\hat{Z}^{(2)}) = \sqrt{v(\hat{Z}^{(2)})} = 82.04 (\text{m}^2).$$

§ 2.4 样本量的确定

2.4.1 确定 n 的一般原则

在抽样调查中, 样本量 n 的确定是一个十分重要的问题. 它不仅与调查的精度有关, 也直接与调查的费用相联系. 注意, 这里的费用含义是广义的, 不仅仅是经费, 也包括涉及的人力与时间等. n 的确定取决于对精度的要求和费用的限制. 对于简单随机抽样, 费用函数甚为简单. 因此在这一节中, 我们主要考虑样本量与精度之间的关系.

例如在例 2.2 居民收入调查中, $n=90$, 最后得到的户平均月收入 Y 的 95% 的置信区间是 (647.67 元, 744.73 元). 这似乎不够精确, 因为置信区间长度较大, 其原因是估计量的方差较大. 为了获得估计的更高精度, 唯一的途径是加大样本量. 但 n 取何值比较适宜, 使它既能满足估计精度要求, 又不会造成浪费呢?

对估计量的精度要求可以以它所允许的最大方差 V (或相应的标准差) 的形式提出来, 或更多的以绝对误差限 (即允许的最大绝对误差) d 或相对误差限 (允许的最大相对误差) r 的形式提出来. 其中 d 与 r 都是在一定概率意义下定义的. 在 1.2.5 段中, 我们已建立了 d 与估计量的方差 (或标准差) 及 r 与估计量的变异系数之间的联系, 即

$$d = u_\alpha \sqrt{V(\hat{\theta})} = u_\alpha S(\hat{\theta}), \quad (2.49)$$

$$r = u_\alpha \frac{S(\hat{\theta})}{\hat{\theta}} = u_\alpha CV(\hat{\theta}). \quad (2.50)$$

由于 $V(\hat{\theta})$ (或 $S(\hat{\theta})$) 是 n 的函数, 由此即可根据 d (或 r) 或给定的估计量最大方差 (或变异系数) 来确定样本量 n 的数值。

2.4.2 总体参数为 \bar{Y} 或 Y 的一般情形

当需要估计的总体参数是总体总和 Y 或平均数 \bar{Y} 时, 所用的基本估计量是样本平均数 \bar{y} 。设 d 是给定置信水平 \bar{y} 的绝对误差限, V 是允许的 \bar{y} 的最大方差, 则根据 y 的方差公式及 (2.49) 式, 有:

$$\begin{aligned}\frac{N-n}{nN} S^2 + \left(\frac{d}{u_\alpha}\right)^2 &= V, \\ (N-n)S^2 &= nN \frac{d^2}{u_\alpha^2}, \\ n\left(\frac{Nd^2}{u_\alpha^2} + S^2\right) &= NS^2,\end{aligned}$$

得

$$n = \frac{NS^2}{N \frac{d^2}{u_\alpha^2} + S^2} = \frac{(u_\alpha S / d)^2}{1 + (u_\alpha S / d)^2 / N}, \quad (2.51)$$

或

$$n = \frac{S^2/V}{1 + S^2/VN}. \quad (2.52)$$

若令

$$n_0 = \left(\frac{u_\alpha S}{d}\right)^2 \text{ 或 } n_0 = \frac{S^2}{V}, \quad (2.53)$$

则

$$n = \frac{n_0}{1 + n_0/N}. \quad (2.54)$$

通常都是由 (2.53) 式计算 n 的一次近似值 n_0 , n_0 比实际需要的 n 要大。若 n_0/N 可以忽略 (例如 $n_0/N < 0.05$), 则就取 n_0 , 否则, 根据 (2.54) 式修正, 得到实际需要的 n 。

例 2.6 在例 2.2 中, 若要求估计的户平均收入的绝对误差在 10 元以内 (置信水平为 95%), 又总体标准差 S 估计为 80 元, 则根据 (2.53) 式, 有

$$n_0 = \left(\frac{1.96 \times 80}{10}\right)^2 \approx 246.$$

由于 $n_0/N = 17.4\%$, 不能忽略, 进而由 (2.54) 式, 得到

$$n = \frac{246}{1 + 246/1328} \approx 233.$$

这就是说, 必须抽取一个样本量不小于 233 的简单随机样本, 才能在 95% 的置信度下保证户月平均收入的估计 \bar{y} 的绝对误差不会超过 10 元。

如果精度的要求是以相对误差限 r 或最大允许的变异系数 C 来表示的, 则从 (2.50) 式出发 (注意, 此时 θ 即为 \bar{Y} , 而 $r\bar{Y}$ 相当于前面的 d), 即可得

$$n_0 = \left(\frac{u_{\alpha} S}{r \bar{Y}} \right)^2 \quad \text{或} \quad n_0 = \frac{1}{C^2} \left(\frac{S}{\bar{Y}} \right)^2, \quad (2.55)$$

其中 S/\bar{Y} 是总体的变异系数。当 n_0/N 不太小时, 也应对 n_0 进行修正, 公式也用 (2.54) 式。

注意, 为了确定样本量, 必须对总体方差 (标准差) 或变异系数事先进行估计 (参见 2.4.4 段的讨论)。

2.4.3 估计总体比例 P 的情形

当欲估计的是总体具有某种特征单元的比例 P 时, 估计量是样本中的相应比例 p 。根据定理 2.6 及 (2.51) 式, 若 d 是置信水平 $1-\alpha$ 下的 p 的绝对误差限, 则

$$n = \frac{\left(\frac{u_{\alpha}}{d} \right)^2 \frac{N}{N-1} PQ}{1 + \left(\frac{u_{\alpha}}{d} \right)^2 \frac{N PQ}{(N-1)N}} = \frac{\frac{u_{\alpha}^2 PQ}{d^2}}{1 + \frac{1}{N} \left(\frac{u_{\alpha}^2 PQ}{d^2} - 1 \right)}. \quad (2.56)$$

若令

$$n_0 = \frac{u_{\alpha}^2 PQ}{d^2}, \quad (2.57)$$

则

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}. \quad (2.58)$$

由于 P 是未知的, 必须事先予以估计。事实上, 当 $0.3 \leq P \leq 0.7$ 时, PQ 很接近于 $P=0.5$ 时的最大值 0.25。在实际问题中, 也往往以 $P=0.5$ 代入 (2.57) 式计算 n_0 。

若对 P 的相对误差提要求, 令 r 是相对误差限 (此时 r 也是 $A = NP$ 估计的相对误差限)。由此可求得

$$n = \frac{\frac{u_{\alpha}^2}{r^2} \cdot \frac{Q}{P}}{1 + \frac{1}{N} \left(\frac{u_{\alpha} Q}{r^2 P} - 1 \right)}. \quad (2.59)$$

于是若对 P 有一个初步的估计 p_0 , 可先计算

$$n_0 = \frac{u_\alpha^2 q_0}{r^2 p_0}, \quad (2.60)$$

其中 $q_0 = 1 - p_0$. 若 n_0/N 不太小, 则根据(2.58)式修正, 得 n . 从(2.59)或(2.60)式可以看到当 P 很小时, 为了达到一定的相对精度, 需要很大的 n .

例 2.7 为估计某县彩色电视机的普及率 P , 用简单随机抽样进行调查. 设允许的 P 的估计最大相对误差为 10% (置信水平取为 95%). 问需调查多少户才能满足要求? (为简单起见, 假定每户至多只拥有一台彩电, $1-f \approx 1$.)

解 这里 $r = 0.1$, $\alpha = 0.05$, $u_\alpha = 1.96$. 若粗略估计该县彩电普及率为 25%, 即 $p_0 = 0.25$, $q_0 = 0.75$. 代入公式(2.60), 有

$$n_0 = \frac{(1.96)^2 \times 0.75}{(0.1)^2 \times 0.25} = 1152.$$

由于 N 很大, 故就取 $n = n_0$.

若实际调查结果为 $p = 0.327$, 则

$$v(p) = \frac{pq}{n-1} = \frac{0.327 \times 0.673}{1151} = 0.0001912,$$

$$s_p = \sqrt{v(p)} = 0.0138.$$

于是该县彩电普及率 P 的 95% 的置信区间为 (30.00%, 35.40%).

当 P 很小 (例如 $P < 0.1$), 即总体中包含所考虑这种特征的单元总数很小时, 如果又没有较好的办法来获得关于 P 的初步估计时, 如何来确定 n 呢?

Haldane(1945) 提出控制估计量变异系数的一种特殊的逆抽样 (inverse sampling). 方法如下: 事先确定一个整数 m ($m > 1$), 进行逐个抽样, 直到抽到 m 个所考虑的特征单元为止. 设 n 是实际的样本量, 则可以证明

$$p' = \frac{m-1}{n-1} \quad (2.61)$$

是 P 的一个无偏估计, 而当 N 很大, $m \geq 10$ 时, 有

$$V(p') \approx \frac{mP^2Q}{(m-1)^2}, \quad (2.62)$$

于是

$$Cv(p') \approx \frac{\sqrt{mQ}}{m-1} < \frac{\sqrt{m}}{m-1}. \quad (2.63)$$

因为 P 很小, 故 $Q \approx 1$, 故上式 $\frac{\sqrt{m}}{m-1}$ 是 $Cv(p')$ 的一个相当接近的上界.

对给定的对 p' 估计的变异系数值, 即可求出 m . 实际所需的样本量 n 是随机的, 但一般都相当大. 因为对 $Cv(p)$ 的一般值, 例如若要求 $Cv(p) < 20\%$, 则 $m \geq 27$, 若要求 $Cv(p) < 10\%$, 则 $m \geq 102$, 考虑到 P 很小的这个事实, n 就相当可观了.

2.4.4 总体方差的预先估计

前面讨论的在调查的设计阶段确定样本量 n 时, 需要对总体的方差 S^2 (或对总体比例 P) 进行预估. 这可以根据以往对类似调查的经验来估计, 或根据对总体结构的了解进行预测. 还有一种常见的情形, 是若在正式调查之前进行试调查, 则可以根据试调查的结果来估计 S^2 (或 P).

如果调查的费用较为昂贵, 必须严格控制 n , 也就需要对 S^2 或 P 作出更可靠的估计. 此时可采用 Stein (1945) 提出的两步抽样 (two step sampling). 在两步抽样中, 第一步先抽 n_1 个单元用来估计 S^2 或 P , 然后确定 n , 第二步再抽其余的 $n - n_1$ 个单元. 下面介绍 D. R. Cox (1952) 在 Stein 工作的基础上提出的在给定精度要求 (V , d 或 C) 下确定 n 的一些结果.

假定 n_1 足够大, 使 $\frac{1}{n_1^2}$ 可以忽略, $n_1 \leq n$, 又 fpc 也可忽略. 记 \bar{y}_1 , s_1^2 或 p_1 是根据第一个样本计算得到的估计量.

1) 假定 Y_i 服从正态分布, 在给定估计量变异系数限 C 时, 估计 \bar{Y} :

$$n = \frac{s_1^2}{C^2 \bar{y}_1^2} \left(1 + 8C^2 + \frac{s_1^2}{n_1 \bar{y}_1^2} + \frac{2}{n_1} \right), \quad (2.64)$$

所得的 \bar{y} 是有偏的, 此时可取 $\bar{Y} = \bar{y}(1 - 2C^2)$.

2) 给定 C , 估计 P :

$$n = \frac{q_1}{C^2 p_1} + \frac{3}{p_1 q_1} + \frac{1}{C^2 p_1 n_1}, \quad (2.65)$$

$$\hat{P} = p - C^2 p/q. \quad (2.66)$$

3) 给定 $V \left(= \frac{d^2}{u_a^2} \right)$, 估计 \bar{Y} :

$$n = \frac{S_1^2}{V} \left(1 + \frac{2}{n_1} \right). \quad (2.67)$$

注意到当 S^2 已知时, $n = \frac{S^2}{V}$, 故 $1 + \frac{2}{n_1}$ 可看作是作两步抽样时, 总

样本量比 S^2 已知时增加的倍数(平均而言).

4) 给定 V , 估计 P :

$$n = \frac{p_1 q_1}{V} + \frac{3}{p_1 q_1} \frac{8 p_1 q_1}{V n_1} + \frac{1}{V n_1} \frac{-3 p_1 q_1}{V n_1}. \quad (2.68)$$

后两项是两步抽样应增加的样本量. 此时 p 也是有偏的, 可取

$$\hat{P} = p + \frac{V(1-2p)}{pq}. \quad (2.69)$$

§ 2.5 放回简单随机抽样

2.5.1 抽样方法及基本特征

前几节讨论的简单随机抽样是不放回抽样, 总体中的任一单元不会在样本中重复出现. 但是在某些实际问题中, 抽样不可能做到完全不放回的. 因此在样本中有可能抽到重复的单元. 例如在对交通车辆或行人的调查中, 当固定在某个路口抽样时就有可能抽到重复的车辆或行人. 又如对影剧院观众的调查以及对野生动物的调查也有类似的情形. 因此有时考虑放回抽样(sampling with replacement)是必要的. 另一个需要考虑放回抽样的原因是: 在放回抽样中, 由于被抽到的单元在下一次抽样前都放回到总体中, 因此每次抽样时总体的结构不变. 因而放回抽样中的每次抽样是相互独立的, 这一点使它的数学处理相对简单得多.

在本节中简单讨论等概率的放回抽样. 具体方法是每次从总体中随机抽取(使总体中每个单元被抽中的概率都相等)一个单元, 经观测记录其指标值后, 放回总体中去, 然后再在总体中随机抽取下一个单元. 这种抽样也称为放回简单随机抽样. 为了研究这种抽样的性质, 先给出以下的引理:

引理 2.3 在大小为 N 的总体中, 按放回简单随机抽样抽取样本量为 n 的一个样本, 用 t_i 表示总体中第 i 个单元在样本中的出现次数 ($t_i = 0, 1, 2, \dots, n; i = 1, 2, \dots, N$), 则

$$E(t_i) = \frac{n}{N} \quad (i = 1, 2, \dots, N), \quad (2.70)$$

$$V(t_i) = \frac{n}{N} \left(1 - \frac{1}{N}\right) \quad (i = 1, 2, \dots, N), \quad (2.71)$$

$$\text{Cov}(t_i, t_j) = -\frac{n}{N^2} \quad (i \neq j). \quad (2.72)$$

证明 由于每次抽样都是随机抽取的, 即总体中每个单元被抽中的

概率都为 $\frac{1}{N}$. 因此 t_i 都服从二项分布 $B\left(n, \frac{1}{N}\right)$, 从而 (2.70)、(2.71) 式成立.

为推导 t_i 与 $t_j (i \neq j)$ 的协方差, 我们计算

$$\begin{aligned} E(t_i t_j) &= \sum_{\substack{k=0 \\ l=0 \\ k+l \leq n}} k \cdot l \cdot \frac{n!}{[k]! [l]! (n-k-l)!} \left(\frac{1}{N}\right)^k \left(\frac{1}{N}\right)^l \left(1 - \frac{2}{N}\right)^{n-k-l} \\ &= \frac{n(n-1)}{N^2} \sum_{\substack{k=0 \\ l=0 \\ k+l \leq n}} \frac{(n-2)!}{(k-1)! (l-1)! (n-k-l)!} \\ &\quad \times \left(\frac{1}{N}\right)^{k-1} \left(\frac{1}{N}\right)^{l-1} \left(1 - \frac{2}{N}\right)^{n-k-l} \\ &= \frac{n(n-1)}{N^2} \left[\frac{1}{N} + \frac{1}{N} + \left(1 - \frac{2}{N}\right) \right]^{n-2} \\ &= \frac{n(n-1)}{N^2}. \end{aligned}$$

从而得

$$\begin{aligned} \text{Cov}(t_i, t_j) &= E(t_i t_j) - E(t_i) E(t_j) \\ &= \frac{n(n-1)}{N^2} - \frac{n^2}{N^2} = -\frac{n}{N^2}. \quad \blacksquare \end{aligned}$$

2.5.2 总体平均数 \bar{Y} 估计量 \bar{y} 的性质

定理 2.9 放回简单随机样本的平均数

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (2.73)$$

是总体平均数 $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$ 的无偏估计, 且

$$V(\bar{y}) = \frac{N-1}{N} \cdot \frac{S^2}{n} = \frac{\sigma^2}{n}, \quad (2.74)$$

其中

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2, \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2. \quad (2.75)$$

证明 1) 对每次抽样, 总体中的任意一个单元 Y_i 都有 $1/N$ 的概率被抽到, 故对每次抽样的结果 y_i , 有

$$E(y_i) = \sum_{i=1}^N \frac{1}{N} \cdot Y_i = \bar{Y}, \quad (2.76)$$

$$V(y_i) = \sum_{i=1}^N \frac{1}{N} \cdot Y_i^2 - \bar{Y}^2 = \sigma^2 \quad (i = 1, 2, \dots, N). \quad (2.77)$$

对不同的 i , y_i 是相互独立的, 因此

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^n E(y_i) = \frac{1}{n} \cdot n\bar{Y} = \bar{Y},$$

$$V(y) = \frac{1}{n^2} \sum_{i=1}^n V(y_i) = \frac{1}{n} \cdot n\sigma^2 = \frac{\sigma^2}{n} = \frac{N-1}{N} \cdot \frac{S^2}{n}.$$

证明 2) 记 t_i 为样本中总体第 i 个单元出现的次数, 于是 \bar{y} 可表示成:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^N t_i Y_i. \quad (2.78)$$

根据引理 2.3 并注意到所有 Y_i 是常数,

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^N Y_i E(t_i) = \frac{1}{n} \cdot \frac{n}{N} \sum_{i=1}^N Y_i = \bar{Y},$$

$$V(y) = \frac{1}{n^2} \left[\sum_{i=1}^N Y_i^2 V(t_i) + 2 \sum_{i < j}^N Y_i Y_j \text{Cov}(t_i, t_j) \right]$$

$$= \frac{1}{n^2} \left[\frac{n(N-1)}{N^2} \sum_{i=1}^N Y_i^2 - \frac{2n}{N^2} \sum_{i < j}^N Y_i Y_j \right]$$

$$= \frac{1}{nN} \left[\frac{N-1}{N} \sum_{i=1}^N Y_i^2 - \frac{2}{N} \sum_{i < j}^N Y_i Y_j \right]$$

$$= \frac{1}{nN} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{\sigma^2}{n}. \blacksquare$$

定理 2.10 对放回简单随机抽样, 样本方差

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2.79)$$

是 σ^2 的无偏估计.

证明 根据 (2.76) 及 (2.77) 式以及定理 2.9, 我们有

$$E \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right] = E \left[\sum_{i=1}^n y_i^2 - n\bar{y}^2 \right] = \sum_{i=1}^n E(y_i^2) - nE(\bar{y}^2)$$

$$= \sum_{i=1}^n \{ V(y_i) + [E(y_i)]^2 \} - n \{ V(\bar{y}) + [E(\bar{y})]^2 \}$$

$$= n(\sigma^2 + \bar{Y}^2) - n \left(\frac{\sigma^2}{n} + \bar{Y}^2 \right) = (n-1)\sigma^2.$$

从而得 $E(s^2) = E \left[\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right] = \sigma^2. \blacksquare$

推论 对放回简单随机抽样, $V(y)$ 的一个无偏估计是

$$v(\hat{y}) = \frac{s^2}{n}. \quad (2.80)$$

定理 2.10 与数理统计中简单样本(独立同分布样本)的性质是一致的. 因为此时总体按通常定义的方差是 σ^2 而不是 S^2 . 样本方差 s^2 并不

是 S^2 的无偏估计, 这是与不放回的简单随机抽样不同的地方, 请读者注意.

2.5.3 设计效应 (deff)

从定理 2.9 可知, 放回简单随机抽样样本均值 \bar{y} 的方差 V_{srswr} 比一般的不放回简单随机抽样样本均值 \bar{y} 的方差 V_{srswor} 要大, 因为两者之比:

$$\frac{V_{srswr}}{V_{srswor}} = \frac{\frac{N-1}{N} \cdot \frac{S^2}{n}}{\frac{N}{N-n} \cdot \frac{S^2}{n}} = \frac{N-1}{N-n} \approx \frac{N}{N-n} = \frac{1}{1-f} > 1.$$

从直观上解释, 这是因为放回抽样有可能重复抽到同一单元, 而同一单元并不会提供更多的信息, 因此放回抽样的效率要比不放回抽样的低.

为比较不同抽样的效率, Kish (1965) 引进一个称为设计效应 (design effect, 简记为 deff) 的量. 它定义为某个特定抽样设计估计量的方差与相同样本量 (不放回) 简单随机抽样的估计量方差之比, 即

$$\text{deff} = \frac{\text{所考虑抽样设计估计量的方差}}{\text{相同样本量 (不放回) 简单随机抽样估计量的方差}}. \quad (2.81)$$

若 $\text{deff} < 1$, 表明所考虑的抽样的效率高于简单随机抽样. 反之, 若 $\text{deff} > 1$, 则它的效率低于简单随机抽样. 对于放回简单随机抽样, 它的 $\text{deff} \approx \frac{1}{1-f}$. 因此, 如果 f 不是太小, 采用放回简单随机抽样是不合算的.

deff 在确定一个复杂抽样设计所需的样本量 n 时有很大的作用. 由于对一定的精度要求, 确定简单随机抽样所需的样本量 n' 比较容易 (§ 2.4). 如果一个复杂抽样的 deff 可以估计, 那么为达到相同的精度要求, 所需的样本量应为

$$n = n'(\text{deff}), \quad (2.82)$$

2.5.4 \bar{Y} 的另一种估计量

前面提到在放回随机抽样中, 由于样本中可能包含重复, 而重复单元并不提供额外的信息. 因此可以将这些重复单元去掉, 考虑另一种估计量. 我们仍以估计总体平均数 \bar{Y} 为例加以说明.

设 y'_1, y'_2, \dots, y'_d 是放回简单随机样本中 d 个不同单元的数值 ($d \leq n$), 注意: 这里仅指不同单元, 而并不排除不同单元有相同的指标值的可能性. 考虑估计量

$$\bar{y}' = \frac{1}{d} \sum_{i=1}^d y'_i, \quad (2.83)$$

它仍是 \bar{Y} 的一个无偏估计. 此外可以证明

$$\begin{aligned} V(\bar{y}') &= \left[E\left(\frac{1}{d}\right) - \frac{1}{N} \right] \frac{N}{N-1} \sigma^2 \\ &\approx \left(\frac{1}{n} - \frac{1}{2N} + \frac{n-1}{12N^2} \right) \frac{N}{N-1} \sigma^2. \end{aligned} \quad (2.84)$$

这里的近似只是省略了 $E\left(\frac{1}{d}\right)$ 展开式中高于 $\frac{1}{N^2}$ 阶的量. 由此可知, $V(\bar{y}')$ 一般要小于 $V(\bar{y})$, 即 \bar{y}' 的精度高于 \bar{y} . $V(\bar{y}')$ 的一个估计是:

$$v(\bar{y}') = \left(\frac{1}{d} - \frac{1}{N} \right) s_d^2, \quad (2.85)$$

其中

$$s_d^2 = \begin{cases} 1, & \text{当 } d=1; \\ \frac{1}{d-1} \sum_{i=1}^d (y'_i - \bar{y}')^2, & \text{当 } d \geq 2. \end{cases} \quad (2.86)$$

§ 2.6 利用随机数骰子和随机数表进行随机抽样的方法

在 2.1.3 段中已提到在实施简单随机抽样时常采用随机数法. 在实际抽样中, 最好使用随机数骰子或现成的随机数表. 本节具体介绍利用随机数骰子或随机数表进行随机抽样的方法. 这些方法不仅适用于通常的(不放回)简单随机抽样、放回简单随机抽样, 也是其他随机抽样(概率抽样)的基础. 例如第 5 章中的各种不等概率抽样在实施时也要采用这里介绍的基本方法.

2.6.1 随机数骰子及其使用方法

随机数骰子是由均匀材料制成的正二十面体(通常的骰子是正六面体, 即正方体), 面上刻有 0~9 的数字各 2 个. 图 2.1 是随机数骰子的底视图与俯视图. 每盒骰子由箱体、盒盖、泡沫塑料垫及若干个(通常是 3~6 个)不同颜色的骰子组成. 使用随机数骰子时可以像普通骰子那样

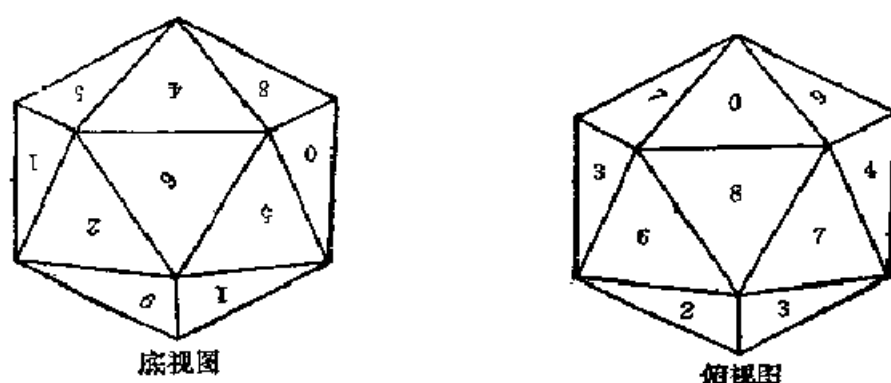


图 2.1 随机数骰子

用掷的方法,但正规的方法是将一个或几个骰子放在盒中,拿去泡沫塑料垫,水平地摇动盒子,使骰子充分旋转,最后打开盒子,读出骰子表示的数字。一个骰子一次产生一个0~9的随机数。要产生一个 m 位数字的随机数,就需要同时使用 m 个骰子(事先规定好每种颜色所代表的位数,例如红色表示百位数,蓝色表示十位数,黄色表示个位数等),或将一个骰子使用 m 次(规定第一次产生的数字为最高位数字,最后一次产生的数字为最末位即个位数等)。特别规定 m 个骰子的数字(或一个骰子 m 次产生的数字)都为0时,表示 10^m 。

当使用随机数骰子进行抽样时,特别是如何根据摇随机数骰子方法获得的随机数 R_0 来读取所要求的随机数 R 有多种方法。下面是我国国家标准GB10111《利用随机数骰子进行随机抽样的方法》中规定的适用于简单随机抽样的读取随机数的方法。

在每种方法中,首先要确定使用的骰子个数(或一个骰子重复摇动的次数) m , m 取决总体大小 N ,且有如下关系:

$$10^{m-1} < N \leq 10^m.$$

记 m 个骰子表示的随机数为 R_0 ,则读取的随机数,也即表示简单随机样本中抽到的总体单元号 R 可用以下三种方法:

方法一 若骰子表示的 $R_0 \leq N$,则取 $R = R_0$;若 $R_0 > N$,则舍弃不用,另行重摇。重复上述过程,直到取得 n 个不同的随机数为止。

例 $N = 725$, $m = 3$. 若 $R_0 = 725, 234, 839$. 则保留前2个,后一个舍弃,重摇。

方法二 如果骰子表示的 $R \leq N$,则取 $R = R_0$;如果 $R_0 > N$,设 $R_0 = K_1 N + R_1$ ($0 \leq R_1 < N$),当 $(K_1 + 1)N > 10^m$ 时,舍弃,重摇。而当 $(K_1 + 1)N \leq 10^m$ 时,取 $R = R_1$ 或 $R = N$ (若 $R_1 = 0$)。重复上述过程,直

到获得 n 个不同的随机数为止。

例 $N = 350$, $m = 3$.

若 $R_0 = 211$, 则取 $R = R_0$.

若 $R_0 = 452 = 1 \times 350 + 102$, $K_1 = 1$, $R_1 = 102$, 且 $(K_1 + 1)N = 2 \times 350 = 700 < 10^3$, 于是取 $R = R_1 = 102$.

若 $R_0 = 810 = 2 \times 350 + 110$, $K_1 = 2$, 由于 $(K_1 + 1)N = 3 \times 350 = 1050 > 10^3$, 故舍弃, 重摇。

方法三 若骰子表示的随机数 $R_0 \leq N$, 则取 $R = R_0$; 若 $R_0 > N$, 则取一个大于 N 的适当整数 M (一般为方便起见取 $M = 2 \times 10^{m-1}$, $2.5 \times 10^{m-1}$, $3 \times 10^{m-1}$ 或 $5 \times 10^{m-1}$ 等). 设 $R_0 = K_2 M + R_2$ (K_2 为整数, $0 \leq R_2 < N$), 则当 $(K_2 + 1)M > 10^m$ 时, 舍弃, 重摇; 当 $(K_2 + 1)M \leq 10^m$ 时, $R = R_2$ 或 $R = N$ (若 $R_2 = 0$). 重复上述过程, 直到获得 n 个不同的随机数为止。

例 $N = 4562$, $m = 4$, 取 $M = 5000$.

若 $R_0 = 3150$, 取 $R = R_0$.

若 $R_0 = 6897 = 1 \times 5000 + 1897$, $K_2 = 1$, $(K_2 + 1)M = 10^4$, $R_2 = 1897$, 故取 $R = R_2 = 1897$.

第二种方法与第三种方法都是为了提高效率, 减少舍弃重摇次数所采取的措施, 尤其是对方法三, 在适当选用 M 时, 既方便又快速。

上述方法也适用于放回简单随机抽样。此时 R 的读取方法也可用上述三种方法的任何一种。所不同的是, 此时连续获得的 n 个随机数 R 即是抽中的样本单元号码, 而不必计较它们是否重复。

2.6.2 随机数表的使用方法

随机数表是将 0 到 9 的数字随机排列而成的。表的产生也有多种方法, 例如反复利用摇动随机数骰子, 将每个骰子表示的数字排列起来就构成随机数表。更多的情况是利用大型计算机, 采用专门设计的程序产生的伪随机数, 产生的伪随机数的循环周期愈长愈好 (至少要求在 10^{10} 之上), 此外, 还应通过各种独立性与随机性的检验。本书末附有五页随机数表, 每页有 $50 \times 50 = 2500$ 个随机数字。排版的方式只是为了使用方便, 在使用时可以根据情况灵活掌握。例如关于排列顺序可以按行从左至右, 到该行结尾时再转下一行, 也可以按列从上至下, 到结尾时再转下一列等等。

在使用时,为克服个人可能有的习惯倾向,增加随机性,首先确定使用的随机数的页数与起点.这也用随机数来确定.譬如说,闭上眼睛将笔放倒在某页随机数表中,以笔尖碰到的数字确定选用的随机数表的页数:例如 0, 1 选用第一页, 2, 3 选用第二页,……, 8, 9 选用第五页等.其次决定随机数的起点,还是闭上眼睛,将笔放倒,笔尖碰到的数字及下一个数字作为起点的行数(必要时减去 25, 50 或 75),用同样的方法再决定列数,这样就决定了起点.

起点确定后,以下的步骤与上一小节介绍的使用随机数骰子的方法相仿,先确定需要的位数 m ,然后按一定的顺序读随机数字,相当于随机数骰子产生的 R_0 ,最后按 GB10111 规定的三种方法中的任何一种确定随机数 R .

第 3 章

分 层 抽 样

§ 3.1 一 般 描 述

3.1.1 定义与记号

定义 3.1 如果大小为 N 的总体分成 L 个不相重迭的子总体, 其大小分别为 N_1, N_2, \dots, N_L (N_h 皆已知, $\sum_{h=1}^L N_h = N$), 每个子总体称为层(stratum). 从每层中独立进行抽样, 这种抽样方法称为分层抽样(stratified sampling), 所得的样本称为分层样本(stratified sample).

在分层抽样中, 若每层的抽样都是简单随机的, 则称为分层随机抽样(stratified random sampling). 所得的样本称为分层随机样本(stratified random sample).

在我国社会经济统计中, 分层抽样有时也称为类型抽样, 这是因为在一些实际问题中, “层”常按照调查对象的不同类型而划分的. 例如在全国性调查中, 将全国各省按经济发达程度或地理位置分层; 在住户家计调查中, 按户主的职业分层; 在对企业调查中, 按企业的行业及规模分层等等.

以后我们用下标 h 表示层的编号 ($h=1, 2, \dots, L$);

用 Y_{hi}, y_{hi} 分别表示总体和样本中关于指标 \mathscr{Y} 的第 h 层第 i 单元的值;

用 $W_h = N_h/N$ 表示层权, 它是已知的;

用 $f_h = n_h/N_h$ 表示 h 层中的抽样比, 其中 n_h 是第 h 层中抽样的样本量.

$$\bar{Y}_h = \sum_{i=1}^{N_h} Y_{hi} / N_h, \quad \bar{y}_h = \sum_{i=1}^{n_h} y_{hi} / n_h$$

分别为 h 层(总体)均值与样本均值;

$$S_h^2 = \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 / (N_h - 1), \quad s_h^2 = \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 / (n_h - 1)$$

分别为 h 层的(层内)方差与样本方差,

3.1.2 分层抽样适用的场合和优点

分层抽样是一种常用的抽样技术, 以下情况都使我们有理由考虑采用分层抽样.

1) 在调查中不仅需要对总体的参数进行估计, 也需要对层的参数进行估计且考虑它们的精度. 例如在一项全国性调查中, 既要求获得全国的结果, 也需要有分省的结果.

2) 使样本更具代表性. 这是因为分层抽样中每层一定有单元被抽到, 从而样本的均匀性更好.

3) 使实施中的组织管理及数据汇总都比较方便. 分层抽样中的数据收集、汇总和处理都可按层独立进行. 如果层是按一定行政系统划分时, 就可按各自的行政系统组织, 而分层样本的数据汇总与处理相当简便.

4) 对不同层可以按照具体情况和条件采用不同的抽样方法. 例如在一些层中用等概率抽样, 而在另一些层中用不等概率抽样; 或者在一些层中用二阶抽样; 在另一些层中需用三阶或四阶抽样.

5) 分层抽样可以提高估计量的精度. 这也是采用分层抽样的原因之一. 在下面的讨论中, 我们将看到在分层抽样中, 层间变差不进入最后估计量的抽样误差中, 因此当层内单元指标差异较小, 而层间差异较大时, 分层抽样的精度就可以有较大幅度的提高.

当然分层抽样也会带来某些技术问题, 首先是层的划分, 有时在实际中分层并不容易, 需要收集必要的资料, 从而耗费额外的费用. 另外, 分层抽样要求各层的大小都是已知的, 当它们不能精确得知时, 就需要通过别的手段进行估计. 这不仅增加了抽样设计的复杂性, 而且也会带进新的误差.

§ 3.2 估计量及其性质

3.2.1 估计量的构造

在分层抽样中, 对总体均值 \bar{Y} 的估计采用各层均值 Y_h 的估计 \hat{Y}_h 按层权 W_h 的加权平均, 即

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^L W_h \hat{\bar{Y}}_h = \frac{1}{N} \sum_{h=1}^L N_h \hat{\bar{Y}}_h. \quad (3.1)$$

特别, 对分层随机抽样, $\hat{\bar{Y}}_h$ 一般取为 h 层的样本均值 \bar{y}_h , 因而 \bar{Y} 用以下简单估计:

$$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h. \quad (3.2)$$

总体总和 Y 的简单估计为

$$\hat{Y}_{st} = N \bar{y}_{st} = \sum_{h=1}^L N_h \bar{y}_h. \quad (3.3)$$

3.2.2 基本性质

定理 3.1 对一般的分层抽样, 若 $\hat{\bar{Y}}_h$ 是 \bar{Y}_h 的无偏估计 ($h=1, 2, \dots, L$), 则 $\hat{\bar{Y}}_{st}$ 是 \bar{Y} 的无偏估计.

证明 $E(\hat{\bar{Y}}_{st}) = E\left(\sum_{h=1}^L W_h \hat{\bar{Y}}_h\right) = \sum_{h=1}^L W_h \bar{Y}_h = \bar{Y}$. ■

定理 3.2 对一般的分层抽样,

$$V(\hat{\bar{Y}}_{st}) = \sum_{h=1}^L W_h^2 V(\hat{\bar{Y}}_h). \quad (3.4)$$

证明 因为各层的抽样是相互独立的, 因此 $\hat{\bar{Y}}_h$ 相互独立, 从而定理得证. ■

定理 3.3 对于分层随机抽样, 作为 \bar{Y} 的简单估计 \bar{y}_{st} , 有

$$E(\bar{y}_{st}) = \bar{Y}, \quad (3.5)$$

$$V(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{n_h} (1 - f_h) \quad (3.6)$$

$$= \sum_{h=1}^L \left(\frac{1}{n_h} - \frac{1}{N_h} \right) W_h^2 S_h^2 \quad (3.7)$$

$$= \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^L \frac{W_h^2 S_h^2}{N_h}. \quad (3.8)$$

证明 从定理 3.2 及定理 2.2 即得. 其中 (3.8) 式中的第二项

$$\sum_{h=1}^L \frac{W_h^2 S_h^2}{N_h} = \sum_{h=1}^L \frac{W_h^2 S_h^2}{N_h}$$

表示考虑有限总体校正因子引起的方差的减少. ■

定理 3.4 对分层随机抽样, $V(\bar{y}_{st})$ 的一个无偏估计是

$$v(\bar{y}_{st}) = \sum_{h=1}^L \frac{W s_{h1}^2}{n_h} (1 - f_h) \quad (3.9)$$

$$= \sum_{h=1}^L \frac{W_h s_h^2}{n_h} - \sum_{h=1}^L \frac{W_h s_h^2}{N}, \quad (3.10)$$

其中

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^L (y_{hi} - \bar{y}_h)^2 \quad (3.11)$$

是第 h 层中所抽样本的样本方差.

证明 根据定理 2.4, s_h^2 是 S_h^2 的无偏估计, 从而由定理 3.3 即得证. ■

$v(y_{st})$ 要求每层的样本量 $n_h \geq 2$. 在 $n_h = 1$ 的情形需作特殊处理, 参见 3.7.3 段.

定理 3.5 对分层随机抽样, Y 的简单估计 $\hat{P}_{st} = N \bar{y}_{st}$ 有如下性质:

$$1^\circ E(\hat{P}_{st}) = Y; \quad (3.12)$$

$$2^\circ V(\hat{P}_{st}) = \sum_{h=1}^L N_h(N_h - n_h) \frac{S_h^2}{n_h}; \quad (3.13)$$

$$3^\circ v(\hat{P}_{st}) = \sum_{h=1}^L N_h(N_h - n_h) s_h^2 / n_h \quad (3.14)$$

是 $V(\hat{P}_{st})$ 的一个无偏估计.

证明 由 \hat{P}_{st} 的定义及定理 3.3、定理 3.4 即可推得. ■

3.2.3 比例分配及自加权样本

在分层抽样中, 一个重要的问题是总的样本量在各层中的分配问题. 这里有两种考虑: 一是出于精度和费用的考虑, 如何分配能使总的精度最高(在一定费用限制下)? 如果对层的估计也有精度要求的话, 还要保证各层的样本量要求. 由于不同层的抽样与调查费用可能有差别, 因此还需要有经济的观点. 另一方面是基于数据处理的考虑, 如何分配能使事后的数据处理比较简洁, 也就是说应尽可能使估计量及其方差估计都有简单的形式, 使数据汇总工作量小, 省时省力.

定义 3.2 在分层抽样中, 若每层的样本量 n_h 都与层的大小 N_h 成比例, 即

$$\frac{n_h}{N_h} = \frac{n}{N} \quad \text{或} \quad f_h = f \quad (h=1, 2, \dots, L), \quad (3.15)$$

则称样本量的这种分配为比例分配 (proportional allocation). (3.15) 式也可写成如下形式:

$$\frac{n_h}{n} = \frac{N_h}{N} = W_h (h = 1, 2, \dots, L), \quad (3.16)$$

比例分配最早是由 Bowley 于 1926 年提出的.

对于比例分配的分层随机抽样, 总体中的任何一个单元, 不管它是哪一层的, 进入样本的概率都为 $f = n/N$. 因此, 比例分配分层随机样本是一种等概率抽取方法(equal probability selection method)形成的样本. 此时作为总体均值 \bar{Y} 的简单估计 y_{st} 等于

$$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h = \sum_{h=1}^L \frac{n_h}{n} \sum_{i=1}^{n_h} y_{hi} = \frac{1}{n} \sum_{h=1}^L \sum_{i=1}^{n_h} y_{hi} = y, \quad (3.17)$$

而总体总和 Y 的估计为:

$$\hat{Y}_{st} = N y_{st} = \frac{N}{n} \sum_{h=1}^L \sum_{i=1}^{n_h} y_{hi} \triangleq \frac{N}{n} y, \quad (3.18)$$

其中

$$y = \sum_{h=1}^L \sum_{i=1}^{n_h} y_{hi} \quad (3.19)$$

是所有样本观测值的总和. 因此, 对比例分配的分层随机样本, 估计量有特别简单的形式.

定义 3.3 对于一种抽样方法, 若总体总和的一个无偏估计可表成最小(基本)样本单元(即个体)观测值总和的一个常数倍, 即

$$\hat{Y} = k y, \quad (3.20)$$

则称这种样本(或估计量)为自加权的(self-weighting)或等加权的(equi-weighting).

当我们用样本观测值来估计总体时, 一种较为自然的估计是将每个样本单元的观测值赋以一个适当的权 ω_i , 然后求和, 也即考虑如下的线性无偏估计:

$$\hat{Y} = \sum_i \omega_i y_i, \quad (3.21)$$

其中 y_i 是样本中最小单元的观测值. 因此对于自加权样本就是意味着所有的权 ω_i 都相等. 等概率抽取方法得到的样本通常是自加权的, 虽然两者从概念上并不完全一致.

由于自加权样本估计量特别简单, 因此只要有可能, 在抽样设计时, 使最终样本为自加权的就可大大简化调查以后的数据处理, 特别是大规模的多指标的调查, 自加权样本的优点尤其明显. 不过, 也应看到, 在大规模抽样调查中, 特别是在涉及多阶抽样中, 要保证最终获得的样本是严格自加权的, 也不是很容易的事. 这里的困难主要不是理论上的(在设计

时要做到这一点并不十分困难), 而是在实际抽样实施时, 常会发生偏离原定设计的情况.

从(3.19)式可以看到, 对于分层随机抽样, 只要做到比例分配, 所得的样本即是自加权的

比例分配分层随机抽样估计量的方差也有比较简单的形式. 事实上, 根据(3.6)、(3.15)及(3.16)式, 此时

$$V_{\text{prop}}(\bar{y}_{st}) = \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2. \quad (3.22)$$

注意: $\sum_{h=1}^L W_h S_h^2$ 是各层层内方差 S_h^2 的按层权的加权平均, 若各层内方差相等或近似相等, 记为 S_w^2 , 则

$$V_{\text{prop}}(\bar{y}_{st}) = \frac{1-f}{n} S_w^2. \quad (3.23)$$

3.2.4 一个简单的实验例子

例 3.1 设总体的 $N=6$, 分成两层, 其单元指标值 Y_{ht} 如表 3.1 所示.

表 3.1 一个简单的分层总体的 Y_{ht} 值

$h \setminus t$	1	2	3
1	0	1	2
2	4	6	11

从表 3.1 中, 易见

$$W_1 = W_2 = 3/6 = 1/2,$$

$$\bar{Y}_1 = 1, \bar{Y}_2 = 7, \bar{Y} = 4, S_1^2 = 1, S_2^2 = \sqrt{13}.$$

考虑在这个总体中抽一个 $n=4$ 比例分配的分层随机样本, 这意味着 $n_1 = n_2 = 2$. 所有可能的样本有 9 个, 对每个样本计算 $\bar{y}_{st}(=\bar{y})$ 及 $\bar{y}_{st} - \bar{Y}$, 结果如表 3.2 所示.

经验证:

$$E(\bar{y}_{st}) = \frac{1}{9} [2.75 + 4.00 + \cdots + 5.00] = \frac{36}{9} = 4 = \bar{Y}.$$

这表明 \bar{y}_{st} 是无偏的. 又

表 3.2 从表 3.1 总体抽取的 $n=4$ 按比例分配的全部可能分层样本

样本编号	$y_{11}, y_{12}; y_{21}, y_{22}$	$\sum_h \sum_i y_{hi}$	y_{st}	$\bar{y}_{st} \quad \bar{Y}$
1	0, 1; 4, 6	11	2.75	-1.25
2	0, 1; 4, 11	16	4.00	0
3	0, 1; 6, 11	18	4.50	0.50
4	0, 2; 4, 6	12	3.00	1.00
5	0, 2; 4, 11	17	4.25	0.25
6	0, 2; 6, 11	19	4.75	0.75
7	1, 2; 4, 6	13	3.25	0.75
8	1, 2; 4, 11	18	4.50	0.50
9	1, 2; 6, 11	20	5.00	1.00
合 计			36.00	0

$$V(\bar{y}_{st}) = \frac{1}{9} [(-1.25)^2 + 0^2 + \cdots + (1.00)^2] = \frac{5.25}{9} = \frac{7}{12}.$$

注意：在应用 $V(\hat{\theta}) = E[\theta - E(\hat{\theta})]^2$ 这个公式直接计算估计量方差时， E 是对所有可能样本求平均的。如果用 (3.22) 式求，结果为

$$V(y_{st}) = \frac{1}{n} f \sum_{h=1}^2 W_h S_h^2 = \frac{1+13}{3 \times 4 \times 2} = \frac{7}{12},$$

与前面的结果一致。

§3.3 最优分配

3.3.1 最优分配的定义

定义 3.4 在分层随机抽样中，对给定费用，使 $V(\bar{y}_{st})$ 达到最小，或对给定的 \bar{y}_{st} 的方差 V 使总费用最小的各层样本量的分配称为最优分配 (optimum allocation)。

在这一节中主要考虑简单的线性费用函数，总费用

$$C = c_0 + \sum_{h=1}^L c_h n_h, \quad (3.24)$$

其中 c_0 是与单元抽取量无关的费用，例如包括组织宣传费用，分层及编制抽样框的费用等。 c_h 是在第 h 层中抽取一个单元的平均费用，包括调查员所费的时间（也包括工资、津贴等）、旅行费用、调查测试费用等。

如果从一个单元至另一单元的调查旅行费用比较昂贵，就可能需要采用稍为复杂的费用函数。例如 Beardwood 等 (1959) 提出以下的费用

函数:

$$c_0 + \sum_h t_h \sqrt{n_h}, \quad (3.25)$$

其中 t_h 是到达每个单元的平均旅行费用.

3.3.2 主要结果

定理 3.6 对分层随机抽样, 若费用函数是简单线性的(3.24), 则最优分配是

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^L W_h S_h / \sqrt{c_h}} = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^L N_h S_h / \sqrt{c_h}} \quad (h=1, 2, \dots, L). \quad (3.26)$$

证明 1 令 $O' = O - c_0 = \sum_{h=1}^L c_h n_h$,

$$V' = V + \sum_{h=1}^L \frac{W_h S_h^2}{N} = \sum_h \frac{W_h^2 S_h^2}{n_h}.$$

则在给定总费用 O 下极小化 \bar{y}_{st} 的方差 V 与在给定 V 下极小化 O 两者都等价于极小化:

$$\begin{aligned} V'O' &= \left(\sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} \right) \left(\sum_{h=1}^L c_h n_h \right) \\ &= \left[\sum_{h=1}^L \left(\frac{W_h S_h}{\sqrt{n_h}} \right)^2 \right] \left[\sum_{h=1}^L (\sqrt{c_h n_h})^2 \right]. \end{aligned} \quad (3.27)$$

根据 Cauchy-Schwarz 不等式, 对任意 $a_h \geq 0$, $b_h \geq 0$, 有

$$\left(\sum_h a_h^2 \right) \left(\sum_h b_h^2 \right) \geq \left(\sum_h a_h b_h \right)^2. \quad (3.28)$$

等号当且仅当

$$\frac{b_h}{a_h} = K = \text{const}$$

时才成立.

于是有
$$V'O' \geq \left(\sum_{h=1}^L W_h S_h \sqrt{c_h} \right)^2.$$

它仅在以下情况达到极小值(上式等号成立):

$$\frac{\sqrt{c_h n_h}}{W_h S_h / \sqrt{n_h}} = \frac{n_h \sqrt{c_h}}{W_h S_h} = K = \text{const} \quad (3.29)$$

上式即意味着

$$n_h = K \frac{W_h S_h}{\sqrt{c_h}} \quad (h=1, 2, \dots, L) \quad (3.30)$$

对所有的 h 求和, 即得到使 $V'O'$ 达到极小的最优分配为:

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^L W_h S_h / \sqrt{c_h}} = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^L W_h S_h / \sqrt{c_h}} \quad (h=1, 2, \dots, L). \quad (3.31)$$

证明 2 用 Lagrange 乘法: 如在总费用 O 固定下, 极小化为 $V(y_{st})$, 则约束条件为

$$O = c_0 + \sum_{h=1}^L c_h n_h. \quad (3.32)$$

$$\begin{aligned} \text{令} \quad T &= V(y_{st}) + \lambda \left(O - c_0 - \sum_{h=1}^L c_h n_h \right) \\ &= \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} - \sum_{h=1}^L \frac{W_h S_h^2}{N} + \lambda \left(O - c_0 - \sum_{h=1}^L c_h n_h \right). \end{aligned}$$

对所有的 h 求 T 对 n_h 的偏导数, 并令其为零, 得:

$$\begin{aligned} -\frac{W_h^2 S_h^2}{n_h^2} + \lambda c_h &= 0, \\ \frac{W_h S_h}{\sqrt{c_h}} &= \sqrt{\lambda} n_h. \end{aligned} \quad (3.33)$$

由于 $\sqrt{\lambda}$ 是常数, 故

$$n_h \propto \frac{W_h S_h}{\sqrt{c_h}}.$$

对所有的 h 求和, 可得到

$$n_h = n \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^L W_h S_h / \sqrt{c_h}} \quad (h=1, 2, \dots, L).$$

在 (3.32) 条件下, 根据 (3.33) 还可解出 n . 事实上, 关于 (3.33) 对所有 h 求和, 可得:

$$\sqrt{\lambda} = \frac{1}{O - c_0} \sum_{h=1}^L \sqrt{c_h} W_h S_h.$$

于是

$$n_h = \frac{O - c_0}{\sqrt{c_h}} \cdot \frac{W_h S_h}{\sum_{h=1}^L \sqrt{c_h} W_h S_h} \quad (h=1, 2, \dots, L). \quad (3.34)$$

定理 3.6 表明 n_h 与 $N_h(W_h)$ 及 S_h 成正比, 与 $\sqrt{c_h}$ 成反比, 这就是说, 层愈大, 层内变差愈大, 而在该层抽样中平均每单元的费用愈小, 则在该层中的抽样应愈多.

3.3.3 Neyman (最优) 分配

如果每层中单位抽样费用相等, 也即 $c_h = c$ 时, 则最优分配简化为:

$$\frac{n_h}{n} = \frac{W_h S_h}{\sum_{h=1}^L W_h S_h} = \frac{N_h S_h}{\sum_h N_h S_h}. \quad (3.35)$$

这个结果早在 1923 年就被俄国学者 Tschuprow 给出, 但由于当时的历史条件, 可惜一直未被人注意到, 直到 1934 年为 Neyman 重新给出. 因此在文献中, 这种简单形式的最优分配常被称为 Neyman 分配.

在 Neyman 分配情形, 所能达到的最小方差为:

$$V_{\min}(\bar{y}_{st}) = \frac{(\sum_h W_h S_h)^2}{n} - \frac{\sum_h W_h S_h^2}{N}. \quad (3.36)$$

这只要将(3.26)式中的 n_h 代入 $V = \sum_h \frac{W_h^2 S_h^2}{n_h}$, 即可得到上式右端的第一项.

例 3.2 对于 3.2.4 段中的实验例子仍取 $n=4$, 但按最优分配 (设每层中的单位抽样费用相同), 则根据(3.35)式, 有

$$n_1 = \frac{n W_1 S_1}{W_1 S_1 + W_2 S_2} = \frac{4 \times 1}{4.6056} = 0.87 \approx 1;$$

$$n_2 = \frac{n W_2 S_2}{W_1 S_1 + W_2 S_2} = \frac{4 \times 3.6056}{4.6056} = 3.13 \approx 3.$$

最优分配的可能样本只有 3 个, 相应的样本单元值及 y_{st} 如表 3.3 所示.

表 3.3 从表 3.1 总体抽取的 $n=4$ 按最优分配的全部可能样本

样 本 编 号	$y_{11}, y_{21}, y_{32}, y_{33}$	y_{st}	$\bar{y}_{st} - \bar{Y}$
1	0, 4, 6, 11	3.5	-0.5
2	1, 4, 6, 11	4.0	0
3	2, 4, 6, 11	4.5	0.5
合 计		12.0	0

$$E(y_{st}) = \frac{1}{3} (3.5 + 4.0 + 4.5) = 4 = \bar{Y}.$$

故 \bar{y}_{st} 仍是无偏的. 按定义直接计算 y_{st} 的方差为:

$$V(y_{st}) = \frac{1}{3} [(-0.5)^2 + 0^2 + (0.5)^2] = 0.167.$$

这个结果与用(3.6)式的结果一致. 但若按(3.36)式, \bar{y}_{st} 的最小方差应为:

$$\begin{aligned}
 V_{\min}(\bar{y}_{st}) &= \frac{1}{n} (\sum W_h S_h)^2 - \frac{1}{N} (\sum W_h S_h^2) \\
 &= \frac{1}{4} \times \frac{1}{4} (1 + 3.6056)^2 - \frac{1}{6} \times \frac{1}{2} (1 + 19) = 0.158.
 \end{aligned}$$

这是理论上能达到的最小值。实际上由于 n_h 只能取整数值，我们在计算时已将计算值 $n_1 = 0.87$, $n_2 = 3.13$ 都归整为 1 与 3，从而实际达到的方差比上述理论值稍大。

3.3.4 某些层需要超过 100% 抽样时的修正

若抽样比 $f = n/N$ 较大，而某些个别层的 S_h 也很大，则按最优分配计算，这些层的 n_h 有可能超过 N_h 。此时可以证明实际最优分配是对这些层进行 100% 抽样的，然后将剩下样本量按最优分配的公式分配。在 Neyman 分配情形，严格的步骤如下（证明留作练习）：

假定 $n_1 > N_1$ ，则令 $\tilde{n}_1 = N_1$ ，

$$\tilde{n}_h = (n - N_1) \frac{W_h S_h}{\sum_{h=2}^L W_h S_h} \quad (h \geq 2). \quad (3.37)$$

若所有的 $\tilde{n}_h \leq N_h$ ($h \geq 2$)，则分配合理，实际配置按 \tilde{n}_h 分配。否则，若有 $\tilde{n}_2 > N_2$ ，则令 $\tilde{n}'_1 = N_1$, $\tilde{n}'_2 = N_2$ ，

$$\tilde{n}'_h = (n - N_1 - N_2) \frac{W_h S_h}{\sum_{h=3}^L W_h S_h} \quad (h \geq 3). \quad (3.38)$$

若所有的 $\tilde{n}'_h \leq N_h$ ($h \geq 3$)，则分配合理。否则，再重复上述过程，直至所有的 $\tilde{n}_h \leq N_h$ 为止。

此时，最优分配达到的（最小）方差公式 (3.36) 也需作相应的修改：

$$V'_{\min}(\bar{y}_{st}) = \frac{1}{n'} (\sum'_h W_h S_h)^2 - \frac{1}{N} \sum'_h W_h S_h^2, \quad (3.39)$$

其中 \sum'_h 为仅对最后实际抽样的 $\tilde{n}_h < N_h$ 的层求和， n' 为这些层中抽样的单元总数。

§ 3.4 分层随机抽样在精度上的得益

3.4.1 与简单随机抽样的比较

在通常情况下，分层随机抽样的精度要比简单随机抽样的高，也即估

计量的方差较小. 由于分层随机抽样的精度与样本量的分配有密切关系, 因此这里不包括明显不合理分配的分层抽样. 事实上, 对任何一个总体, 都可设计一种特别的分配, 使分层随机抽样的精度比简单随机抽样的还要差. 当然这没有任何意义.

在这一小节中, 我们将最优分配、比例分配分层随机抽样与相同样本量的简单随机抽样作精度的比较.

定理 3.7 若 $\frac{1}{N_h} \ll 1$ ($h = 1, 2, \dots, L$), 则最优分配(Neyman 分配情形)分层随机抽样估计量 \bar{y}_{st} 的方差 V_{opt} 、比例分配分层随机抽样 \bar{y}_{st} 的方差 V_{prop} 与简单随机抽样 y 的方差 V_{sts} 之间有如下关系:

$$V_{opt} \leq V_{prop} \leq V_{sts}. \quad (3.40)$$

证明 根据最优分配的定义, $V_{opt} \leq V_{prop}$. 故只需证明

$$V_{prop} = \frac{1}{n} f \sum_{h=1}^L W_h S_h^2 \leq \frac{1-f}{n} S^2 = V_{sts}.$$

考虑总体各单元指标值 Y_M 对总体均值 \bar{Y} 离差平方和的分解:

$$\begin{aligned} (N-1)S^2 &= \sum_{h=1}^L \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y})^2 \\ &= \sum_{h=1}^L \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 + \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2 \\ &= \sum_{h=1}^L (N_h - 1) S_h^2 + \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2. \end{aligned} \quad (3.41)$$

两端同时除以 $N-1$, 由于对所有的 h , $1/N_h \ll 1$, 故

$$\frac{N_h}{N-1} \approx \frac{N_h}{N} \approx W_h.$$

从而得

$$S^2 \approx \sum_{h=1}^L W_h S_h^2 + \sum_{h=1}^L W_h (\bar{Y}_h - \bar{Y})^2. \quad (3.42)$$

上式第二项即是层间平方和, 是非负的, 因此有

$$\begin{aligned} V_{sts} &\approx V_{prop} + \frac{1-f}{n} \sum_{h=1}^L W_h (\bar{Y}_h - \bar{Y})^2, \\ V_{prop} &\leq V_{sts}. \end{aligned} \quad (3.43)$$

从而定理得证. ■

3.4.2 何时分层及最优分配的精度得益最大

现在我们考虑最优分配与比例分配分层随机抽样方差的差,

$$V_{\text{prop}} - V_{\text{opt}} = \frac{1}{n} \left[\sum_{h=1}^L W_h S_h^2 - \left(\sum_{h=1}^L W_h S_h \right)^2 \right] \\ - \frac{1}{n} \sum_{h=1}^L W_h (S_h - \bar{S})^2, \quad (3.44)$$

其中 $S = \sum_h W_h S_h$

是 S_h 按 W_h 的加权平均.

于是根据(3.43)与(3.41)式, 当 $\frac{1}{N_h} \ll 1$ 时, 有

$$V_{\text{int}} \approx V_{\text{opt}} + \frac{1}{n} \sum_{h=1}^L W_h (S_h - \bar{S})^2 + \frac{1-f}{n} \sum_{h=1}^L W_h (\bar{Y}_h - \bar{Y})^2. \quad (3.45)$$

上式右端的第二项是各层标准差的差异, 它可通过考虑最优分配得以消除, 而第三项是各层均值的差异, 它可通过比例分配的分层抽样得以消除. (3.45)式也表明当各层均值差异愈大, 则一般的分层(以比例分配为其代表)的效益愈高, 而当各层的标准差相差较大时, 最优分配又可比比例分配有较大的得益.

最理想的分层是按调查指标 \mathscr{Y} 的数值分. 由于 Y_i 本身是调查对象, 不能直接利用, 但可以通过与 \mathscr{Y} 相关的一个或多个辅助变量来分层.

对于最优分配, 还需用关于层的标准差 S_h 的信息, 因此需要事先进行估计, 譬如说根据以往的调查指标或与它相关的辅助指标的信息. 也可以用与 S_h 直接有联系的量. 譬如说, 如果层内变差系数不大, 则可用与 $W_h \bar{Y}_h$ 也即与 Y_h 成正比的分配形式. 另一种情况是利用层内的极差, 也即用与 $W_h r_h$ (r_h 为 h 层的极差)成比例的分配形式. 这些都可以看成是最优分配的一些变通方法. 由于 S_h 实际上不知需要估计, 加上其他一些原因, 因此最优分配的实际得益并没有公式表示的那么大(关于偏离最优分配造成的影响详见3.4.6段). 相对而言, 由于比例分配的样本是自加权的, 且一般而言(除非 S_h 差别过于悬殊)距最优分配并不太远, 故更受实际工作者所欢迎. 通常, 若比例分配的方差仅比(理论上的)最优分配的方差大10%~20%, 则用比例分配仍是值得的.

3.4.3 分层随机抽样精度反比简单随机抽样差的情形

理论上并不排除出现分层随机抽样的效果反比简单随机抽样差的情况(正如前面已指出的, 这里不包括人为的不合理分配), 虽然在合理分层情形, 这是不大可能发生的.

根据(3.41)式, 有

$$\begin{aligned}
V_{\text{str}} &= \frac{1-f}{n} S^2 = \frac{1-f}{n} \sum_{h=1}^L \frac{N_h-1}{N-1} S_h^2 + \frac{1-f}{n(N-1)} \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2 \\
&= \frac{1-f}{n} \sum_{h=1}^L \frac{1}{N(N-1)} [N_h(N-1) - (N-N_h)] S_h^2 \\
&\quad + \frac{1-f}{n(N-1)} \sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2 \\
&= \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 - \frac{1-f}{n(N-1)} \frac{1}{N} \sum_{h=1}^L (N-N_h) S_h^2 \\
&\quad + \frac{1-f}{n(N-1)} \sum_{h=1}^L N_h (Y_h - \bar{Y})^2 \\
&= V_{\text{prop}} + \frac{1-f}{n(N-1)} \left[\sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2 - \frac{1}{N} \sum_h (N-N_h) S_h^2 \right].
\end{aligned} \tag{3.46}$$

上面推导过程中未作任何近似, 因此如果

$$\sum_{h=1}^L N_h (\bar{Y}_h - \bar{Y})^2 < \frac{1}{N} \sum_{h=1}^L (N-N_h) S_h^2, \tag{3.47}$$

就有 $V_{\text{str}} < V_{\text{prop}}$ 的情况出现. 而 (3.47) 式成立的情况是存在的. 为简单起见, 设 $S_h^2 = S_w^2$ (此时最优分配与比例分配等价), (3.47) 式右端即为

$$\frac{S_w^2}{N} \sum_{h=1}^L (N-N_h) = \frac{S_w^2}{N} (LN - N) = (L-1) S_w^2.$$

因为

$$S_b^2 \triangleq \frac{\sum_h N_h (\bar{Y}_h - \bar{Y})^2}{L-1} \tag{3.48}$$

即是层间方差, 因此 (3.47) 等价于

$$S_b^2 < S_w^2.$$

这也就是对 Y_{hi} 作方差分析时 $F < 1$ 的情形. 这种情况是不难列举的

例 3.3 一个 $N=15$, $L=3$ 的总体如表 3.4 所示.

表 3.4

$h \backslash i$	1	2	3	4	5	\bar{Y}_h	S_h^2
1	3	8	9	4	6	6	6.5
2	0	2	4	6	8	4	10
3	3	7	5	9	1	5	10

经平方和分解得到的方差分析表如表 3.4' 所示。

表 3.4'

变差来源	平方和	自由度	方差(均方)	F
层间	$\sum_h N_h (\bar{Y}_h - \bar{Y})^2 = 10$	$L - 1 = 2$	$S_b^2 = 5.00$	0.57
层内	$\sum_h \sum_i (Y_{hi} - \bar{Y}_h)^2 = 106$	$L(N_h - 1) = 12$	$S_{cr}^2 = 8.83$	
总计	$\sum_h \sum_i (Y_{hi} - \bar{Y})^2 = 116$	$N - 1 = 14$	$S^2 = 8.29$	

对这个总体, 无论哪种分配的分层随机抽样的效果都比简单随机抽样的差。其根源是对这个总体的分层不合理(平均层内方差大于总体方差)。

3.4.4 从样本估计分层随机抽样精度的得益

前面的讨论都是在已知总体及各层具体结构情况下进行的。实际上有关总体的精确参数是未知的。现在的问题是在一个分层随机抽样实施以后, 能否根据样本数据来估计由于分层获得的精度上的好处? 或估计这个分层抽样的设计效应?

根据定理 3.4, \bar{y}_{st} 的方差可以从样本中获得估计, 因此问题的焦点是如何用一个分层样本来对相同样本量下简单随机抽样的方差 $V(\bar{y})$ 进行估计。J. N. K. Rao (1962) 给出了如下的定理:

定理 3.8 根据分层随机样本, 同样样本量的简单随机样本对总体均值估计的方差 $V(\bar{y})$ 的一个无偏估计为:

$$v_{\text{ers}} = \frac{N-n}{n(N-1)} \left[\frac{1}{N} \sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}^2 - \bar{y}_{st}^2 + v(\bar{y}_{st}) \right], \quad (3.49)$$

其中 $v(\bar{y}_{st})$ 由 (3.10) 给出。

证明 由定理 2.2, $V(\bar{y})$ 可改写为:

$$V_{\text{ers}} \triangleq V(\bar{y}) = \frac{N-n}{nN} S^2 = \frac{N-n}{n(N-1)} \left[\frac{1}{N} \sum_{h=1}^L \sum_{i=1}^{N_h} Y_{hi}^2 - \bar{Y}^2 \right].$$

因为
$$E \left(\sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}^2 \right) = \sum_{h=1}^L \frac{N_h}{n_h} \frac{n_h}{N_h} \sum_{i=1}^{N_h} Y_{hi}^2 = \sum_{h=1}^L \sum_{i=1}^{N_h} Y_{hi}^2,$$

$$E[v(y_{st})] = V(\bar{y}_{st}) = E(\bar{y}_{st}^2) - \bar{Y}^2,$$

从而
$$E[y_{st}^2 - v(y_{st})] = \bar{Y}^2.$$

于是
$$E(v_{\text{ers}}) = V_{\text{ers}}. \quad \blacksquare$$

根据定理 3.8, 即可计算一个分层随机抽样的设计效应 deff:

$$\widehat{\text{deff}} = \frac{v(\bar{y}_{st})}{v_{\text{srn}}}. \quad (3.50)$$

(3.49) 式中

$$\frac{1}{N} \sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}^2 = \sum_{h=1}^L W_h \left(\frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}^2 \right)$$

也可以用层样本方差 s_h^2 与均值 \bar{y}_h 来表示, 从而得到

$$v_{\text{srn}} = \frac{N}{n(N-1)} \left[\sum_{h=1}^L W_h s_h^2 + \sum_h \frac{W_h s_h^2}{n_h} + \sum_{h=1}^L W_h \bar{y}_h^2 - \bar{y}_{st}^2 + v(\bar{y}_{st}) \right]. \quad (3.51)$$

当 n_h 都比较大, 例如 $n_h > 50$ 时 (此时 n 更大), $\sum_h \frac{W_h s_h^2}{n_h}$ 与 $v(\bar{y}_{st})$ 的值可忽略不计, 此时有

$$v_{\text{srn}} \approx \frac{N}{nN} \left[\sum_h W_h s_h^2 + \sum_h W_h \bar{y}_h^2 - \bar{y}_{st}^2 \right]. \quad (3.52)$$

对比例分配, (3.51) 可简化为:

$$v_{\text{srn}} = \frac{N-n}{n(N-1)} \left[\frac{n-1}{n} s^2 + v(\bar{y}_{st}) \right] \quad (3.53)$$

其中

$$(n-1)s^2 = \sum_{h=1}^L \sum_{i=1}^{n_h} (y_{hi} - \bar{y})^2$$

即为样本离差平方和 (此时 $\bar{y}_{st} = \bar{y}$), 若 n 足够大, 有近似公式:

$$v_{\text{srn}} \approx \frac{1}{n} f s^2. \quad (3.54)$$

3.4.5 数值例子——关于职工月平均奖金额的调查

例 3.4 为调查某市企业、机关与事业单位职工的月平均奖金, 将职工所属单位按性质分成 4 层, 有关数据如表 3.5 所示. 其中层内标准差 S_h 系估计数值.

表 3.5 职工月平均奖金调查按所属单位分层情况

h (层名称)	N_h (职工人数)	W_h (层权)	S_h (层内标准差)	$W_h S_h$
1 全民企业	15220	0.54164	14	7.58296
2 集体企业	8710	0.30996	23	7.12908
3 合资企业	850	0.03025	44	1.33100
4 机关事业单位	3320	0.11815	5	0.59075
合 计	28100			16.63379

给定样本总量 $n=600$, 按指定分配 $n_1=300$, $n_2=200$, $n_3=n_4=50$ 以及比例分配、最优分配(Neyman 分配)进行分层随机抽样. 实际分配的样本量数值列于表 3.6.

表 3.6 职工月平均奖金调查样本量的分配情况

h	指定分配	比例分配	最优分配
1	300	325	274
2	200	186	257
3	50	18	48
4	50	71	21

调查后经初步计算各层样本均值与方差的数据见表 3.7.

表 3.7 职工月平均奖金调查各层样本均值 \bar{y}_h 与方差 s_h^2

h	指定分配		比例分配		最优分配	
	\bar{y}_h	s_h^2	\bar{y}_h	s_h^2	\bar{y}_h	s_h^2
1	25.50	196.70	28.10	180.78	26.7	186.50
2	35.40	379.16	34.20	453.20	36.80	382.18
3	48.50	1340.81	42.30	895.20	46.90	1128.27
4	14.20	18.57	15.30	24.98	15.80	21.3

分别对三种不同的分配计算 y_{st} 及 $v(y_{st})$, 并对各自的 deff 进行估计:

解 1) 指定分配

$$\begin{aligned}
 y_{st} &= \sum_{h=1}^4 W_h \bar{y}_h \\
 &= 0.54164 \times 25.50 + 0.30996 \times 35.40 + 0.03025 \times 48.50 \\
 &\quad + 0.11815 \times 14.20 = 27.9293,
 \end{aligned}$$

$$\begin{aligned}
 v(\bar{y}_{st}) &= \sum_{h=1}^4 W_h^2 \frac{s_h^2}{n_h} = \sum_{h=1}^4 W_h^2 \frac{s_h^2}{N_h} \\
 &= 0.4042 + 0.0095 = 0.3947,
 \end{aligned}$$

$$\begin{aligned}
 v_{\text{BRS}} &= \frac{N-n}{n(N-1)} \left[\sum_{h=1}^4 W_h s_h^2 - \sum_h \frac{W_h^2 s_h^2}{n_h} + \sum_h W_h \bar{y}_h^2 - \bar{y}_{st}^2 + v(\bar{y}_{st}) \right] \\
 &= 0.0016311 (266.8034 - 0.4042 + 835.6102 \\
 &\quad - 780.0458 + 0.3947) \\
 &= 0.0016311 \times 322.3589 = 0.5258,
 \end{aligned}$$

$$\widehat{\text{deff}} = \frac{v(y_{st})}{v_{\text{BRS}}} = \frac{0.3947}{0.5258} = 0.75.$$

2) 比例分配

$$\bar{y}_{st} = \sum_h W_h y_h = 28.9080,$$

$$v(\bar{y}_{st}) = \frac{1}{n} \sum_h W_h s_h^2 = 0.4378,$$

$$v_{\text{srst}} = \frac{N}{n(N-1)} \left[\sum_h W_h s_h^2 - \sum_h \frac{W_h s_h^2}{n_h} + \sum_h W_h y_h^2 - \bar{y}_{st}^2 + v(\bar{y}_{st}) \right] \\ = 0.0016311 \times 302.5952 = 0.4936,$$

$$\widehat{\text{deff}} = \frac{v(\bar{y}_{st})}{v_{\text{srst}}} = \frac{0.4378}{0.4936} = 0.89.$$

这里计算 v_{srst} 仍用一般情形的(3.51)式, 而没有用(3.53)式. 根据表 3.7 给出的数据, 若用(3.53)式, 则所有样本数据的离差平方和应按下列式计算:

$$(n-1)s^2 = \sum_h (n_h-1)s_h^2 + \sum_h n_h(y_h - \bar{y}_{st})^2.$$

3) 最优分配

$$\bar{y}_{st} = \sum_h W_h y_h = 29.1538,$$

$$v(\bar{y}_{st}) = \sum_{h=1}^4 \frac{W_h s_h^2}{n_h} - \sum_{h=1}^4 \frac{W_h s_h^2}{N} \\ = 0.3778 - 0.0091 = 0.3687,$$

$$v_{\text{srst}} = \frac{N}{n(N-1)} \left[\sum_h W_h s_h^2 - \sum_h \frac{W_h s_h^2}{n_h} + \sum_h W_h y_h^2 - \bar{y}_{st}^2 + v(\bar{y}_{st}) \right] \\ = 0.0016311 \times 308.0257 = 0.5024,$$

$$\widehat{\text{deff}} = \frac{v(\bar{y}_{st})}{v_{\text{srst}}} = \frac{0.3687}{0.5024} = 0.73.$$

注意: 在本例中, 为了比较, 用三种分配抽取了二个分层随机样本, 上面对同样样本量($n=600$)简单随机抽样方差的估计 v_{srst} 是对三个样本分别计算的. 鉴于本例的实际情况, 当然还可使用别的估计方法. 譬如将三个估计联合起来, 或者用(3.45)式进行估计, 因为此时我们对 V_{opt} 及层标准差 S_h 及均值 \bar{Y}_h 都可得到较为精确的估计. 有兴趣的读者可作一尝试.

最后我们指出, 在本例中比例分配的精度还不及指定分配的精度, 原因是这个指定分配已是相当接近最优分配. 而这个“最优”分配事实上也不是真正的最优, 从各层样本方差中可看到, 原先对层方差的估计, 特别是对集体企业($h=2$)及合资企业($h=3$)两层, 是过份了.

3.4.6 偏离最优分配时对方差的影响

在实际问题中, 由于层标准差 S_h 需要估计, n_h 又只能取整等原因, 在考虑最优分配 (以 Neyman 情形为例) 时, 实际所用的样本量分配 \hat{n}_h 与理论上的最优分配 n'_h 会有所偏离. 在这一小节中, 我们讨论由于这种偏离造成的估计量 \bar{y}_{st} 方差的变化.

按实际分配的样本量 \hat{n}_h , 根据 (3.8) 式, 估计量 \bar{y}_{st} 的方差为:

$$V(\bar{y}_{st}) = \sum_h \frac{W_h^2 S_h^2}{\hat{n}_h} - \sum_h \frac{W_h S_h^2}{N}.$$

而理论上最优分配 n'_h 所能达到的最小方差为:

$$V_{\min}(\bar{y}_{st}) = \frac{1}{n} \left(\sum_h W_h S_h \right)^2 - \sum_h \frac{W_h S_h^2}{N}.$$

实际分配的方差与最小方差比较, 方差增加量为

$$V(\bar{y}_{st}) - V_{\min}(\bar{y}_{st}) = \sum_h \frac{W_h^2 S_h^2}{\hat{n}_h} - \frac{1}{n} \left(\sum_h W_h S_h \right)^2. \quad (3.55)$$

根据 (3.35) 式解得

$$W_h S_h = \frac{n'_h}{n} \left(\sum_h W_h S_h \right),$$

代入 (3.55) 式, 有

$$\begin{aligned} V(\bar{y}_{st}) - V_{\min}(\bar{y}_{st}) &= \sum_h \frac{n_h'^2 (\sum_h W_h S_h)^2}{n^2 \hat{n}_h} - \frac{1}{n} \left(\sum_h W_h S_h \right)^2 \\ &= \frac{1}{n^2} \left(\sum_h W_h S_h \right)^2 \left[\sum_h \frac{n_h'^2}{\hat{n}_h} - 2n + n \right] \\ &= \frac{1}{n^2} \sum_h W_h S_h)^2 \sum_h \left(\frac{n_h'^2}{\hat{n}_h} - 2n'_h + \hat{n}_h \right) \\ &= \frac{1}{n^2} \left(\sum_h W_h S_h \right)^2 \sum_h \frac{(\hat{n}_h - n'_h)^2}{\hat{n}_h} \end{aligned} \quad (3.56)$$

$$\triangleq \frac{1}{n^2} \left(\sum_h W_h S_h \right)^2 \sum_h \hat{n}_h g_h^2, \quad (3.57)$$

其中

$$g_h = \frac{\hat{n}_h - n'_h}{\hat{n}_h} \quad (3.58)$$

是 h 层实际样本量与最优分配样本量的相对偏离, 若忽略 fpc, 有

$$V_{\min}(\bar{y}_{st}) \approx \frac{\left(\sum_h W_h S_h \right)^2}{n}.$$

因而此时方差的相对增加为:

$$\frac{V(\bar{y}_{st}) - V_{\min}(\bar{y}_{st})}{V_{\min}(\bar{y}_{st})} \approx \sum_k \frac{\hat{n}_k}{n} g_k^2. \quad (3.59)$$

上式右边是 g_k^2 的加权平均, 它的上限为:

$$g^2 \triangleq \max\{g_k^2\}. \quad (3.60)$$

例 3.5 表 3.8 给出了一个实际分配 \hat{n}_h 与理论最优分配 n'_h 偏离程度的数值及方差相对增加量的计算步骤.

表 3.8 偏离最优分配方差相对增加量的计算

层 (h)	n'_h	\hat{n}_h	$g_h = \frac{ \hat{n}_h - n'_h }{\hat{n}_h}$	$\hat{n}_h g_h^2 = \frac{(\hat{n}_h - n'_h)^2}{\hat{n}_h}$
1	200	180	0.111	2.222
2	150	130	0.154	3.077
3	75	100	0.250	6.250
4	35	50	$g=0.300$	4.500
总 计	460	460		16.049

方差的相对增加量为 $16.049/460 = 3.49\%$, $g = \max\{g_h\} = 0.3$. 因此即使用上限 g^2 , 也仅 9%. 例 3.4 中指定分配与最优分配实际精度比较也说明了这个问题. 结论是: 在最优分配中, n'_h 即使有些误差, 对实际方差影响也不会很大.

3.4.7 多指标情形样本量的分配

关于多指标的调查, 对某个指标的最优分配通常也不会是其他指标的最优的或近似最优的分配. 此时, 最简单的办法是采用比例分配, 不仅因为它形式简单(包括其后的数据处理), 而且对各指标大多能获得颇为满意的结果.

本小节仍从最优分配的角度考虑多指标情形样本量的分配方法. 这些方法本质上都是对不同指标最优分配的某种程度的折衷.

一、各指标最优分配平均法

在众多的指标中, 选取最重要的 k 个, 对每个指标 j 计算最优分配的层样本量 n'_{jh} , 然后求其平均值:

$$n_h = \frac{1}{k} \sum_{j=1}^k n'_{jh} \quad (h=1, 2, \dots, L). \quad (3.61)$$

由于指标之间一般具有一定的相关性, 因此, 各指标的最优分配不会

过于悬殊。取平均后，差别更小。考虑到在计算最优分配时还受到各指标层标准差估计误差的影响，因此在实际问题中， n_h 一般已能满足要求。

二、Chatterjee(1967)方法

设 n'_{jh} 是按第 j 个指标的最优分配，考虑实际分配样本量 n_h 对每个指标偏离其最优分配引起的方差的相对增加 RV_j ，根据(3.59)与(3.58)式，有

$$RV_j \triangleq \frac{V_j(y_{st}) - V_{j, \text{min}}(y_{st})}{V_{j, \text{min}}(y_{st})} \approx \frac{1}{n} \sum_h \frac{(n_h - n'_{jh})^2}{n_h} \quad (j=1, 2, \dots, k), \quad (3.62)$$

取使极小化 RV_j 的平均值

$$\frac{1}{k} \sum_{j=1}^k RV_j \quad (3.63)$$

的 n_h ，结果为：

$$n_h = n \frac{\sqrt{\sum_j n_{jh}'^2}}{\sum_h \sqrt{\sum_j n_{jh}'^2}}. \quad (3.64)$$

Chatterjee 方法的结果(3.64)与平均法结果(3.61)相差甚微。

三、Yates 方法 I(1960)

考虑损失函数

$$\begin{aligned} L &= \sum_{j=1}^k a_j V(y_{j, st}) = \sum_{j=1}^k a_j \sum_{h=1}^r W_h^2 S_{jh}^2 \left(\frac{1}{n_h} - \frac{1}{N} \right) \\ &= \sum_h \frac{W_h^2}{n_h} \left(\sum_j a_j S_{jh}^2 \right) - \frac{1}{N} \sum_h W_h \left(\sum_j a_j S_{jh}^2 \right) \\ &\triangleq \sum_h \frac{W_h^2}{n_h} \left(\sum_j a_j S_{jh}^2 \right) - L_0. \end{aligned} \quad (3.65)$$

若费用函数仍是简单线性的形式 $O = c_0 + \sum_h c_h n_h$ ，极小化

$$(O - c_0)(L - L_0) = \left(\sum_h c_h n_h \right) \left(\sum_h \frac{W_h^2}{n_h} \sum_j a_j S_{jh}^2 \right). \quad (3.66)$$

根据 Cauchy Schwarz 不等式(3.28)，极小值当且仅当

$$\frac{\sqrt{c_h n_h}}{W_h \sqrt{\sum_j a_j S_{jh}^2}} = \frac{\sqrt{c_h n}}{W_h \sqrt{\sum_j a_j S_{jh}^2}} = K = \text{const} \quad (3.67)$$

时达到。记

$$A_h \triangleq \sqrt{\sum_j a_j S_{jh}^2}, \quad (3.68)$$

则最优分配为

$$n_h \propto W_h A_h / \sqrt{c_h}. \quad (3.69)$$

从而

$$n_h = \frac{n W_h A_h / \sqrt{c_h}}{\sum_h (W_h A_h / \sqrt{c_h})}. \quad (3.70)$$

四、Yates 方法 II(1960)

对每个指标, 给定要求的精度 V_j , 即要求

$$\sum_{h=1}^L \frac{W_h^2 S_{jh}^2}{n_h} - \sum_{h=1}^L \frac{W_h S_{jh}^2}{N} \leq V_j, \quad (j=1, 2, \dots, k) \quad (3.71)$$

在约束条件(3.71)及

$$0 \leq n_h \leq N_h \quad (h=1, 2, \dots, L) \quad (3.72)$$

之下, 极小化

$$C = c_0 + \sum_h c_h n_h$$

即可求得 n_h , 从而化为一个线性规划问题.

Booth 与 Sedransk(1969)将上述问题化成方法 I 处理, 从而避免了复杂的计算, 将损失函数定义为

$$V^* = \sum_{j=1}^k \alpha_j V_j. \quad (3.73)$$

取 α_j 与 V_j 成反比. 例如当 $k=2$ 时,

$$\alpha_1 = \frac{V_2}{V_1 + V_2}, \quad \alpha_2 = \frac{V_1}{V_1 + V_2}, \quad V^* = \frac{2V_1 V_2}{V_1 + V_2}. \quad (3.74)$$

§ 3.5 样本总量 n 的确定

3.5.1 估计的总体参数为 \bar{Y} 的情形

当估计的总体参数为总体均值 \bar{Y} 时, 估计量为 \bar{y}_{st} , 设它的允许的最大方差为 V (或规定的绝对误差限为 d , $d^2 = V u_a^2$), 对某种已确定的样本量分配:

$$n_h = w_h n \quad (h=1, 2, \dots, L). \quad (3.75)$$

根据定理 3.3, \bar{y}_{st} 的预期方差为

$$V(\bar{y}_{st}) = \frac{1}{n} \sum_{h=1}^L \frac{W_h^2 S_h^2}{w_h} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2.$$

于是对给定的 V ,

$$n \geq \frac{\sum_{h=1}^L W_h^2 S_h^2 / w_h}{V + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}. \quad (3.76)$$

可取 n 的一次近似值为

$$n_0 = \frac{1}{V} \sum_{h=1}^L \frac{W_h^2 S_h^2}{w_h}. \quad (3.77)$$

若 n_0/N 不能忽略, 则进一步计算

$$n = \frac{n_0}{1 + \frac{1}{NV} \sum_{h=1}^L W_h S_h^2}. \quad (3.78)$$

特别对比例分配的情形, $w_h = W_h$, 代入(3.76)式, 得

$$n = \frac{\sum_{h=1}^L W_h S_h^2}{V + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}. \quad (3.79)$$

此时若先计算

$$n_0 = \frac{1}{V} \sum_{h=1}^L W_h S_h^2, \quad (3.80)$$

则

$$n = \frac{n_0}{1 + \frac{n_0}{N}}. \quad (3.81)$$

对 Neyman 最优分配, $w_h = \frac{W_h S_h}{\sum_h W_h S_h}$, 代入(3.76)式, 得

$$n = \frac{\left(\sum_{h=1}^L W_h S_h \right)^2}{V + \frac{1}{N} \sum_h W_h S_h^2}. \quad (3.82)$$

在需要考虑各层费用不同的情形, 对于简单的线性费用函数及相应的最优分配形式, 当 V 给定时, 从

$$\begin{aligned} V &= \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h} = \sum_{h=1}^L \frac{W_h S_h^2}{N} \\ &= \sum_h \frac{W_h^2 S_h^2}{n} \cdot \frac{\sum_h W_h S_h / \sqrt{c_h}}{W_h S_h / \sqrt{c_h}} = \sum_h \frac{W_h S_h^2}{N} \\ &= \frac{1}{n} \left(\sum_h W_h S_h \sqrt{c_h} \right) \left(\sum_h W_h S_h / \sqrt{c_h} \right) = \sum_h \frac{W_h S_h^2}{N}, \end{aligned}$$

由此可解出

$$n = \frac{\left(\sum_h W_h S_h \sqrt{c_h} \right) \left(\sum_h W_h S_h / \sqrt{c_h} \right)}{V + \sum_h W_h S_h^2 / N}. \quad (3.83)$$

而当总费用 O 给出时, 由于此时最优分配的每层样本量已由(3.34)

式给出, 从而总样本量

$$n = \sum_k n_k = (O - c_0) \frac{\sum_k W_k S_k / \sqrt{c_k}}{\sum_k W_k S_k \sqrt{c_k}}. \quad (3.84)$$

例 3.6 考虑如表 3.9 所示的 $L=2$ 的分层总体:

表 3.9

h	W_h	S_h	c_h
1	0.4	10	4
2	0.6	20	9

设费用函数 $O = \sum_h c_h n_h$, 求使 $V(\bar{y}_{st}) = 1$ 所需要的按最优分配的 n_1 与 n_2 (取 $fpc = 1$).

$$\text{解 } \sum_{h=1}^2 \frac{W_h S_h}{\sqrt{c_h}} = \frac{0.4 \times 10}{2} + \frac{0.6 \times 20}{3} = 2 + 4 = 6.$$

故最优分配为

$$\frac{n_1}{n} = \frac{2}{6} = \frac{1}{3}, \quad \frac{n_2}{n} = \frac{4}{6} = \frac{2}{3}.$$

由 (3.83) 式, 当 $fpc = 1$ 时, $\sum_h W_h S_h^2 / N$ 可忽略不计:

$$n \approx \frac{1}{V} \left(\sum_h W_h S_h c_h \right) \left(\sum_h W_h S_h / \sqrt{c_h} \right).$$

根据题意, $V = 1$. 故

$$n = (0.4 \times 10 \times 2 + 0.6 \times 20 \times 3) \times 6 = 44 \times 6 = 264.$$

从而

$$n_1 = 88, \quad n_2 = 176.$$

此时总费用为

$$O = 88 \times 4 + 176 \times 9 = 1936.$$

若现场实际调查费用 $c'_1 = 2$, $c'_2 = 12$. 则为达到原先要求的 $V = 1$, 调查费用必须适当增加. 若按原样本量分配, 则调查实际费用为

$$O' = 88 \times 2 + 176 \times 12 = 2288.$$

若重新进行最优分配, 此时需要的最小费用可计算如下:

最优分配时, $\left(V + \sum_h \frac{W_h S_h^2}{N} \right) (O - c_0)$ 达到极小值 $(\sum_h W_h S_h \sqrt{c_h})^2$.

在本例中 $c_0 = 0$, $\sum_h \frac{W_h S_h^2}{N}$ 可忽略不计, $V = 1$, 故需要的最小费用为:

$$C'' = (\sum_k W_k S_k \sqrt{c_k})^2 = (0.4 \times 10 \times \sqrt{2} + 0.6 \times 20 \times \sqrt{12})^2 = 2230.$$

这相当于新的最优分配 $n'_1 = 134$, $n'_2 = 164$ [由(3.34)式], 按 n'_1 、 n'_2 计算出的实际费用(2236)比理论上的最小费用(2230)稍大, 这是由于在计算 n'_1 、 n'_2 过程中最后需要取整数造成的.

3.5.2 估计的总体参数为 Y 的情形

当需要估计的总体参数为总体总和 Y 时, 估计量为 $\hat{Y}_{st} = N\bar{y}_{st}$, 设 \bar{V} 是 \hat{Y}_{st} 允许的最大方差, 则将 $V = \bar{V}/N^2$ 代入 3.5.1 段中的有关公式, 即可得到需要的结果. 以下仅对主要情形列出相应公式.

对给定的分配形式($n_h = nw_h$)有

$$n = \frac{\sum_k N_k^2 S_k^2 / w_k}{V + \sum_k N_k S_k^2}, \quad (3.85)$$

$$n_0 = \frac{1}{\bar{V}} \sum_k \frac{N_k^2 S_k^2}{w_k}, \quad n = \frac{n_0}{1 + \frac{1}{\bar{V}} \sum_k N_k S_k^2}. \quad (3.86)$$

Neyman 最优分配:

$$n = \frac{(\sum_k N_k S_k)^2}{V + \sum_k N_k S_k^2}, \quad (3.87)$$

$$n_0 = \frac{1}{\bar{V}} (\sum_k N_k S_k)^2, \quad n = \frac{n_0}{1 + \frac{1}{\bar{V}} \sum_k N_k S_k^2}. \quad (3.88)$$

比例分配:

$$n = \frac{N \sum_k N_k S_k^2}{V + \sum_k N_k S_k^2}, \quad (3.89)$$

$$n_0 = \frac{N}{\bar{V}} \sum_k N_k S_k^2, \quad n = \frac{n_0}{1 + \frac{n_0}{N}}. \quad (3.90)$$

§ 3.6 对总体比例(百分率)的分层随机抽样

前面几节的结果都可以直接用于对总体比例(或百分率) P 估计的分层抽样. 在这一节中, 仅列出主要结果以便于使用. 其中 $P_h = A_h/N_h$, $p_h = a_h/n_h$, 其他符号与前几节相同.

3.6.1 估计量及其方差

总体比例 P 的分层估计为:

$$p_{st} = \sum_h W_h p_h. \quad (3.91)$$

它是 P 的无偏估计. 将 $S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h$ 代入(3.6)式, 即有

$$V(p_{st}) = \frac{1}{N^2} \sum_h \frac{N_h^2 (N_h - n_h)}{N_h - 1} \cdot \frac{P_h Q_h}{n_h} \quad (3.92)$$

$$\approx \sum_h \frac{N_h^2}{N^2} \cdot \frac{N_h - n_h}{N_h} \cdot \frac{P_h Q_h}{n_h} \quad (N_h \approx N_h - 1)$$

$$= \sum_h \frac{W_h^2 P_h Q_h}{n_h} (1 - f_h). \quad (3.93)$$

对比例分配情形:

$$p_{st} = p = \frac{1}{n} \sum_h a_h \quad (3.94)$$

$$V(p_{st}) = \frac{1-f}{nN} \sum_h \frac{N_h^2 P_h Q_h}{N_h - 1} \quad (3.95)$$

$$\approx \frac{1-f}{n} \sum_h W_h P_h Q_h.$$

当用样本数据估计上述的 $V(p_{st})$ 时, 可将 $\frac{P_h Q_h}{n_h - 1}$ 代替上面公式中的 $\frac{P_h Q_h}{N_h - 1}$, 所得到的结果是 $V(p_{st})$ 的无偏估计.

3.6.2 最优分配

对于简单的线性费用函数

$$C = c_0 + \sum_h c_h n_h,$$

最优分配满足

$$n_h \propto N_h \sqrt{\frac{N_h}{N_h - 1}} \sqrt{\frac{P_h Q_h}{c_h}} \approx N_h \sqrt{P_h Q_h / c_h}, \quad (3.96)$$

从而

$$n_h = n \frac{N_h \sqrt{P_h Q_h / c_h}}{\sum_h N_h \sqrt{P_h Q_h / c_h}}. \quad (3.97)$$

3.6.3 分层和最优分配精度上的得益

鉴于在实际问题中, 不同层的 P_h 以及 $P_h Q_h$ 一般不可能相差很大

(特别是 $P_h Q_h$), 因此, 根据 § 3.4 的讨论, 对总体比例的估计, 考虑分层以及最优分配在精度上的得益不会十分显著. 当然, 由于其他考虑, 在这种情形, 分层仍是常被采用的技术.

下面是对一个虚拟的总体来说明对于比例估计的分层随机抽样(比例分配情形)与相同样本量的简单随机抽样精度上的比较. 其中总体分为三层, 层权都为 $1/3$.

表 3.10 一个虚拟的总体分层随机抽样与简单随机抽样比例估计的精度比较

P_h	简单随机抽样 $\frac{n}{1-f} V(p) = PQ$	分层随机抽样(比例分配) $\frac{n}{1-f} V(p_{st}) = \frac{1}{3} \sum P_h Q_h$	设计效应 deff	分层抽样相对精度 n/deff
0.4, 0.5, 0.6	0.25	0.2433	0.97	1.03
0.3, 0.5, 0.7	0.25	0.2233	0.89	1.12
0.2, 0.5, 0.8	0.25	0.1900	0.76	1.32
0.1, 0.5, 0.9	0.25	0.1433	0.57	1.74

表 3.10 中考虑几种 P_h 不同的取值, 总体 P 都为 0.5. 我们看到当各层的 P_h 在 0.3~0.7 之间时, 按比例分配的分层随机抽样相对于简单随机抽样精度提高不多. 后两种情况, 精度虽有较为显著的提高, 但在实际问题中又不大可能出现这种情况.

我们再对另一个虚拟的总体来比较最优分配与比例分配的精度. 总体由两层组成, $W_1 = W_2 = 1/2$, 其中第一层的 $P_1 = 0.5$, 表 3.11 给出了不同的 P_2 值最优分配估计量的方差 V_{opt} 与比例分配估计量方差 V_{prop} 的比.

表 3.11 一个虚拟的总体最优分配与比例分配对比例估计的精度比较

P_2	0.4 (0.6)	0.3 (0.7)	0.2 (0.8)	0.1 (0.9)	0.05 (0.95)
$V_{\text{opt}}/V_{\text{prop}}$	1.00	0.998	0.988	0.941	0.866

从表 3.11 中可见到, 只有当 $P_2 < 0.1$ (或 $P_2 > 0.9$) 时, 最优分配才有较为显著的得益.

不过, 当考虑的比例在各层中的数值 P_h 很小时, P_h 数值本身(特别是它的相对变化)和 $P_h Q_h$ 值都可能有一定的变化(例如在 0.05~0.001 范围内变动), 此时采取分层以及分层中的最优分配则是值得的.

3.6.4 样本量的估计

设 V 是估计量 p_{st} 的最大允许方差, 则当 N_h 都比较大, 使

$$N_h \approx N_h - 1$$

时, 总样本量 n 可按以下公式估计:

比例分配情形:

$$n_0 = \sum_h W_h P_h Q_h / V, \quad (3.98)$$

若 n_0/N 不能忽略,

$$n = \frac{n_0}{1 + \frac{n_0}{N}}. \quad (3.99)$$

最优分配情形:

$$n_0 = (\sum_h W_h \sqrt{P_h Q_h})^2 / V, \quad (3.100)$$

若 n_0/N 不能忽略,

$$n = \frac{n_0}{1 + \frac{1}{NV} \sum_h W_h P_h Q_h}. \quad (3.101)$$

在计算时, 都需要对 P_h 作预先的估计.

§ 3.7 分层技术的充分利用

3.7.1 层的构造

分层抽样的一个主要优点是估计量的精度较高. 为了充分利用分层在精度上的得益, 需要考虑如何来构造层. 例如一项全国性的调查, 如果对省(直辖市、自治区)需要抽样, 那么为提高精度, 应将全国 30 个省(市、自治区)进行分层. 这里的分层实际上即是分类. 我们可以将全国所有省(市、自治区)按经济、文化发达的程度进行分类, 例如利用聚类分析方法根据多种指标将它们分类, 以类作为层.

在本小节中, 我们着重考虑当按一个指标分层时, 层的构造方法. 前已提到, 此时最有效的方法是按调查指标 \mathscr{Y} 的数值分, 需要确定的是层间的分点.

设总体需分成 L 层, 其中 y_0 、 y_L 分别是 \mathscr{Y} 的最小与最大可能值, 设 $y_1 < y_2 < \cdots < y_{L-1}$ 是确定层的 $L-1$ 个分点, 我们的目标是在各层样本量分配原则已定的情况下, 如何确定 $y_1, y_2, \cdots, y_{L-1}$ 的值, 使估计量的方差 $V(\bar{y}_{st})$ 达到极小. $y_1, y_2, \cdots, y_{L-1}$ 称为层的最优分点 (optimum points of stratification).

我们首先讨论比例分配情形. 根据(3.22)式, 此时估计量的方差为:

$$V_{\text{prop}} = \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2.$$

于是最优分点的确定等价于极小化:

$$\begin{aligned} \sum_{h=1}^L W_h S_h^2 &\propto \sum_{h=1}^L N_h S_h^2 \approx \sum_{h=1}^L \sum_{i=1}^{N_h} (Y_{hi} - Y_h)^2 \\ &= \sum_{h=1}^L \left(\sum_{i=1}^{N_h} Y_{hi}^2 - N_h Y_h^2 \right), \end{aligned}$$

式中“ \approx ”是在所有 N_h 都比较大时具有的性质. 由于 $\sum_{h=1}^L \sum_{i=1}^{N_h} Y_{hi}^2$ 是常数, 故最优分点的确定等价于极大化

$$\sum_{h=1}^L N_h \bar{Y}_h^2. \quad (3.102)$$

为求 y_h , 假定其余分点都固定, 注意到在(3.102)式中 y_h 仅对 \bar{Y}_h^2/N_h 与 \bar{Y}_{h+1}^2/N_{h+1} 两项有影响. 由于 y_h 是最优分点, 故以下两式成立:

$$\frac{\bar{Y}_h^2}{N_h} + \frac{\bar{Y}_{h+1}^2}{N_{h+1}} \geq \frac{(\bar{Y}_h - y'_h)^2}{N_h - 1} + \frac{(\bar{Y}_{h+1} + y'_h)^2}{N_{h+1} + 1}, \quad (3.103)$$

$$\frac{\bar{Y}_h^2}{N_h} + \frac{\bar{Y}_{h+1}^2}{N_{h+1}} \geq \frac{(\bar{Y}_h + y''_h)^2}{N_h + 1} + \frac{(\bar{Y}_{h+1} - y''_h)^2}{N_{h+1} - 1}. \quad (3.104)$$

其中 y'_h 与 y''_h 是紧挨着 y_h 的两个单元的值. 当 N_h 和 N_{h+1} 都比较大时, (3.103)与(3.104)两式分别可简化为:

$$(\bar{Y}_{h+1} - \bar{Y}_h)(\bar{Y}_{h+1} + \bar{Y}_h - 2y'_h) \geq 0,$$

$$(\bar{Y}_{h+1} - \bar{Y}_h)(\bar{Y}_{h+1} + \bar{Y}_h - 2y''_h) \geq 0,$$

或
$$y'_h \leq \frac{1}{2}(\bar{Y}_h + \bar{Y}_{h+1}), \quad y''_h \geq \frac{1}{2}(\bar{Y}_h + \bar{Y}_{h+1}).$$

因此

$$y_h \approx \frac{1}{2}(\bar{Y}_h + \bar{Y}_{h+1}) \quad (h=1, 2, \dots, L-1). \quad (3.105)$$

当各层样本量的分配是 Neyman 最优分配时, 根据(3.96)式, 估计量的方差为:

$$V_{\min} = \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^L W_h S_h^2.$$

若 N_h 都很大, 从而 $1/N_h$, $1/N$ 都可忽略情形, 与前面类似的证明方法, 可得最优分点 y_h 满足

$$\frac{S_h^2 + (y_h - \bar{Y}_h)^2}{S_h} \approx \frac{S_{h+1}^2 + (y_h - \bar{Y}_{h+1})^2}{S_{h+1}} \quad (h=1, 2, \dots, L-1). \quad (3.106)$$

从(3.105)与(3.106)可见, 最优分点 y_1, y_2, \dots, y_{L-1} 是一组方程的解. 由于 \bar{Y}_h, S_h 都与 y_h 有关, 因此要解这组方程是十分困难的, 必须通过复杂的迭代才能实现.

为此, 许多作者提出一些近似但快速的求解法. 下面是 Dalenius 与 Hodges(1959)提出的方法.

对于 Neyman 最优分配情形, 目标是使 $\sum_h W_h S_h$ 极小化. 设 \mathscr{Y} 的频数分布为 $f(y)$. 在给定的层中, 将 $f(y)$ 近似看作为常数, 即是服从均匀分布的, 此时

$$W_h \approx f_h(y_h - y_{h-1}), \quad (3.107)$$

$$S_h = \frac{1}{\sqrt{12}}(y_h - y_{h-1}), \quad (3.108)$$

$$\sqrt{12} \sum_{h=1}^L W_h S_h \approx \sum_{h=1}^L f_h \cdot (y_h - y_{h-1})^2 \approx \sum_{h=1}^L (Z_h - Z_{h-1})^2. \quad (3.109)$$

其中 Z_h 是直至 y_h 的 $\sqrt{f(y)}$ 累积和:

$$Z_h = \int_{y_0}^{y_h} \sqrt{f(t)} dt. \quad (3.110)$$

容易证明当 $Z_h - Z_{h-1}$ 都相等时, (3.109) 右端达到极小. 由此可见, 只要 $f(y)$ 已知, 就可按 $\sqrt{f(y)}$ 的累积值来确定分点: 选择这样的 y_h , 使累积 $\sqrt{f(y)}$ 等分即可. 故这种方法称为累积 $f^{1/2}(y)$ 法.

例 3.7 表 3.12 是某地区工薪阶层户月人均收入的频数分布, 欲将它分为 $L=7$ 层, 求层的最优分点.

根据表 3.12 最右列, 累积 \sqrt{f} 值为 683.4557, 欲将它 7 等分, 间距应分 $683.4557/7 = 97.6365$. 于是得 6 个理论最优分点的累积 \sqrt{f} 值及最接近的(实际)分点如表 3.13 所示.

Singh (1971) 建议用累积 $f^{1/3}$ (法). 在一定假定下, 可以证明对一般的累积 $f^{1/\alpha}$ 法 ($2 \leq \alpha \leq 3$), 方差具有 $O(L^{-2})$ 的收敛速度, 比较合理. 在 $\alpha \in [2, 3]$ 中, 保守地看, 取 $\alpha = 1/3$ 最佳.

上述讨论是以调查指标 \mathscr{Y} 的分布为基础的, 在实际情形由于 y 未知, 这个假定是不现实的. 通常以与 \mathscr{Y} 线性密切相关的另一辅助变量 \mathscr{X} 的分布代替 \mathscr{Y} . 例如在车辆运输量的调查中, \mathscr{Y} 是需调查的运量或周转量, 则可取 \mathscr{X} 为车辆的吨位(载货汽车情形)或客位(载客汽车情形)的分布来确定层的最优分点.

另一个与此相关的问题是层数 L 取多大合适? 从精度而言, 当然是 L 取大些为好, 但这势必增大工作量. 因此 L 的确定既与当 L 增大时方

表 8.12 用累积 \sqrt{f} 法确定层的最优点

组 序 号	月人均收入 y^* (元)	$f(y)$	$\sqrt{f(y)}$	累积 $\sqrt{f(y)}^*$
1	100~150	7	2.645751	2.645751
2	150~200	20	4.472136	7.117887
3	200~250	78	8.831761	15.94965
4	250~300	156	12.49000	28.43964
5	300~350	232	15.23155	43.67119
6	350~400	350	18.70829	62.37948
7	400~450	378	19.44222	81.82170
8	450~500	507	22.51666	104.3384
9	500~550	735	27.11088	131.4492
10	550~600	891	29.84962	161.2989
11	600~650	1260	35.49648	196.7953
12	650~700	1674	40.91455	237.7099
13	700~750	1864	43.17407	280.8840
14	750~800	2027	45.02222	325.9062
15	800~850	1907	43.66921	369.5754
16	850~900	1780	42.19006	411.7654
17	900~950	1560	39.49684	451.2623
18	950~1000	1132	33.64521	484.9075
19	1000~1100	1502	38.75564	523.6631
20	1100~1200	843	29.03446	552.6976
21	1200~1300	433	20.80865	573.5062
22	1300~1400	352	18.76166	592.2678
23	1400~1500	153	12.36932	604.6371
24	1500~1600	67	8.185353	612.8224
25	1600~1700	38	6.164414	618.9868
26	1700~1800	9	3.000000	621.9868
27	1800~1900	11	3.316625	625.3034

表 3.13

理论最优分点的累积 \sqrt{f} 值	相应的实际分点
97.6465	500
195.2731	650
292.9096	750
390.5461	900
488.1826	1000
585.8192	1200

*) 注: 表中的分组的组距不完全相等, 前 18 组(人均月收入在 1000 元以内的)以 50 元为组距; 后 9 组(月人均收入超过 1000 元的)以 100 元为组距(分组都是左开右闭的, 即每组包括右端点, 不包括左端点)。后者是前者的两倍。因此, 从第 19 组开始是按 $\sqrt{2f}$ 累计, 相当于将后面的每组拆成两组, 而频数 f 为平均分配的结果。

差的减少速度有关, 又与当 L 增大时费用的增加有关. 通常在二者之间取一个平衡值. 当然也可用恰当的模型描述, 这里就不详细论述了.

3.7.2 多重分层

当调查指标 \mathcal{Y} 与两个或多个辅助变量 $\mathcal{X}_1, \mathcal{X}_2, \dots$ 都线性相关时, 为充分利用分层的效益, 就需要按每个辅助变量分层. 例如在进行家庭调查时, 可按家庭居住的地区、户主的年龄、职业、文化程度等多种指标分层. 此时我们一般的做法是先按最主要的一个变量分成大层, 在大层中再按第二个变量分成子层, 从而引成交叉分层. 当存在多个分层变量时, 这种分层方法即称为多重分层 (multiple stratification).

在多重分层中, 样本量的分配可将每种方式的分层按某种原则分配, 然后再将不同方式分层的分配结果按一定原则进行折衷. 不过最简单也最常用的方法是按每一子层大小成比例的原则进行分配. 以按两个分层变量也即两种方式分层为例, 若按第一种方式分层, 共分成 R 大层, 每一大层的层权为 $W_{.k} (k=1, 2, \dots, R)$; 按第二种方式分层共分成 O 大层, 每一大层的层权为 $W_{.l} (l=1, 2, \dots, O)$, 则每个子层的层权为 $W_{kl} = W_{.k} W_{.l} (k=1, 2, \dots, R; l=1, 2, \dots, O)$. 设总样本量为 n , 则 kl 子层的样本量 $n_{kl} = nW_{kl}$.

在多重分层中, 由于子层总数比较大, 而受费用等因素的限制, n 又不能取得很大时, 就常会出现不能保证每个子层都能分配到样本单元的情形. 仍以两种方式分层为例, 若 $n < RO$, 但 $n \geq \max(R, O)$, 我们可用实验设计的思想来分配样本量. 下面是一个说明性的例子.

例 3.8 某城镇进行货车运输量的抽样调查. 分层原则一是按货车的吨位大小分, 二是按车辆的所属部门及营业性质分. 前者共分为 $R=6$ 层, 后者分为 $O=5$ 层, $RO=30$. 若 $n=9$, 如何来确定各子层的样本量?

首先我们将每个子层及行层 (按第一种方式分层形成的大层) 与列层 (按第二种方式分层形成的大层) 的大小 $N_{kl}, N_{.k}, N_{.l}$ 列成表 3.14 的形式, 其中车辆总数, 即总体大小 $N=977$. 然后我们计算各子层、行层与列层大小对总体大小的比例 $P_{kl}=N_{kl}/N, P_{.k}=N_{.k}/N, P_{.l}=N_{.l}/N$, 再计算 $n=9$ 时按比例分配原则分配给各行层与列层的样本量 $n'_{.k}=nP_{.k}, n'_{.l}=nP_{.l}$. 经过舍入取整为 $n_{.k}$ 与 $n_{.l}$. 上述数据都列在表 3.15 中.

在确定了每个行层及列层的样本量以后, 如何将它们进一步分到子层中去呢? 我们的原则是每个子层被分配到一个样本单元的概率为

表 3.14 某城镇货车按两种方式分层,各子层的大小 N_{hi}

$h \backslash i$	1	2	3	4	5	N_{hi}
1	32	44	40	85	16	217
2	10	8	21	12	13	64
3	24	14	27	54	8	127
4	54	13	35	18	32	152
5	48	26	148	24	42	288
6	65	15	28	9	12	129
$N_{h\cdot}$	283	120	299	202	123	$N=977$

表 3.15 某城镇货车按两种方式分层,各子层在总体中的比例及行层、列层样本量的分配

$h \backslash i$	1	2	3	4	5	P_{hi}	n'_{hi}	n_{hi}
1	0.0328	0.044	0.0409	0.087	0.0164	0.2221	2	2
2	0.0102	0.0082	0.0215	0.0123	0.0133	0.0655	0.59	1
3	0.0246	0.0143	0.0276	0.0553	0.0082	0.1300	1.17	1
4	0.0553	0.0133	0.0358	0.0184	0.0328	0.1656	1.40	1
5	0.0491	0.0266	0.1515	0.0246	0.043	0.2948	2.65	3
6	0.0665	0.0154	0.0287	0.0092	0.0123	0.1320	1.19	1
$P_{h\cdot}$	0.2385	0.1228	0.306	0.2068	0.1259			
$n'_{h\cdot}$	2.15	1.11	2.75	1.86	1.13			
$n_{h\cdot}$	2	1	3	2	1			$n=9$

$n_{hi}, n_{hi}/n^2$, 为实现这一点, 借助一个 $n \times n$ 的方阵, 这个方阵的前 $n_{1\cdot}$ 行对应于第 1 个行层, 第 $n_{1\cdot}+1$ 行至 $n_{1\cdot}+n_{2\cdot}$ 行的 $n_{2\cdot}$ 行对应于第 2 个行层, ……; 将各列也与列层相对应, 先在第一行中随机地抽取一列, 然后在第二行中在其余 $n-1$ 列中随机地抽取一列, 依此类推, 结果是在方阵中每行都有一个格子被抽中, 每列也有一个格子被抽中, 抽中的每一格子所在的行列子层就分配到一个样本单元, 本例中的一次抽取结果如图 3.1 所示(图中的虚线是 $n \times n$ 阶方阵中格子界线, 实线是子层的界线)。

熟悉试验设计的读者立即会想到, 上述在 $n \times n$ 方阵中抽取 n 个格子的过程相当于在一个 n 阶拉丁方中随机地抽一个字母(或数字), 该字母所占有的 n 个格子即是所需要的, 当然这个拉丁方本身字母的排列应该是经过随机化的, 这个概念可以推广到一般的多重分层, 例如按三种或

$h \backslash l$	1	2	3	4	5	$n_{h\cdot}$
1				×		2
2			×			1
3	×					1
4					×	1
5		×	×			3
6		×				1
$n_{\cdot l}$	2	1	3	2	1	9

图 3.1

更多种方式分层的类似步骤可借助 n 阶正交拉丁方来完成。

在确定了在哪些子层内需要抽取样本后, 具体方法则是在子层内进行随机抽取(如果分配的样本量不止一个, 即可用简单随机抽样)。此时一个在 hl 子层内抽取的样本单元被抽到的概率与 $P_{hl}/n_{h\cdot}n_{\cdot l}$ 成比例, 因而是等概率的, 不过当每个 $P_{hl} \approx n_{h\cdot}n_{\cdot l}/n^2$ 时, 其概率是近似相等的。只有在此时, 样本均值可作为总体均值的估计。否则, 应按不等概率抽样(参见第 5 章)方法处理。至于方差估计, 只有当 $n \geq 2 \max(R, C)$ 时, 也即在每个行层与列层中至少有两个样本单元时才能进行。

3.7.3 每层只抽一个单元时的方差估计

在 3.7.2 段讨论了由于多重分层, 层数很多而样本量较小时的抽样问题。在本段讨论的问题与上述也有一定联系, 假定在每一(子)层中只抽取一个单元。在此情形, 我们仍可用 3.2.2 段中的一般公式来估计总体目标量。但此时方差估计就不能按那里的方法进行了, 因为对于 $n_h = 1$, 无法计算层内的样本方差 s_h^2 。但我们可以用层间估计量的差异来估计方差。为方便起见, 我们仅对总体总和的估计 $\hat{P}_{..}$ 的方差进行讨论。

在抽样前将所有层分成两两一组或数层一组。在层数 $L = 2G$ ——即为偶数的情形, 以两层一组为宜, 共分 G 组。设第 j 组的两个样本观测值为 y_{j1}, y_{j2} , 则两层总和的估计分别为:

$$\hat{P}_{j1} = N_{j1}y_{j1}, \quad \hat{P}_{j2} = N_{j2}y_{j2}, \quad (3.111)$$

其中 N_{j1} 与 N_{j2} 分别是这两层的大小, 令

$$v(\hat{P}_{st}) = \sum_{j=1}^G (\hat{P}_{j1} - \hat{P}_{j2})^2, \quad (3.112)$$

将它作为 $V(\hat{P}_{st})$ 的一个估计, 为求它的均值, 考虑将 \hat{P}_{j1} \hat{P}_{j2} 表成:

$$\hat{P}_{j1} - \hat{P}_{j2} = (Y_{j1} - Y_{j2}) + (\hat{P}_{j1} - Y_{j1}) - (\hat{P}_{j2} - Y_{j2}),$$

平方后再求均值, 所有的交叉乘积项皆为零, 于是有

$$\begin{aligned} E(\hat{P}_{j1} - \hat{P}_{j2})^2 &= (Y_{j1} - Y_{j2})^2 + N_{j1}^2 E(y_{j1} - \bar{Y}_{j1})^2 + N_{j2}^2 E(y_{j2} - \bar{Y}_{j2})^2 \\ &= (Y_{j1} - Y_{j2})^2 + N_{j1}(N_{j1} - 1)S_{j1}^2 + N_{j2}(N_{j2} - 1)S_{j2}^2, \end{aligned}$$

故

$$\begin{aligned} E[v(\hat{P}_{st})] &= \sum_{j=1}^G (Y_{j1} - Y_{j2})^2 + \sum_{j=1}^G [N_{j1}(N_{j1} - 1)S_{j1}^2 + N_{j2}(N_{j2} - 1)S_{j2}^2] \\ &= \sum_{j=1}^G (Y_{j1} - Y_{j2})^2 + \sum_{h=1}^L N_h(N_h - 1)S_h^2 \\ &= \sum_{j=1}^G (Y_{j1} - Y_{j2})^2 + V(\hat{P}_{st}). \end{aligned} \quad (3.113)$$

上式表明, 作为 $V(\hat{P}_{st})$ 的估计, $v(\hat{P}_{st})$ 是有偏的, 偏倚为 $\sum_{j=1}^G (Y_{j1} - Y_{j2})^2$. 因此我们在分组时, 应将层和估计相差不多的层作为一组, 以尽可能减少偏倚.

当 L 不为偶数时, 就必须考虑每组层数 $L_j \geq 2$ 的一般情况. 此时 $V(\hat{P}_{st})$ 的估计可取为:

$$v'(\hat{P}_{st}) = \sum_{j=1}^G \frac{L_j}{L_j - 1} \sum_{k=1}^{L_j} \left(\hat{P}_{jk} - \sum_{k=1}^{L_j} \hat{P}_{jk} / L_j \right)^2. \quad (3.114)$$

可以证明

$$E[v'(\hat{P}_{st})] = V(\hat{P}_{st}) + \sum_{j=1}^G \frac{L_j}{L_j - 1} \sum_{k=1}^{L_j} \left(Y_{jk} - \sum_{k=1}^{L_j} Y_{jk} / L_j \right)^2. \quad (3.115)$$

因此分组原则仍然是同组内的层和愈接近愈好. 当所有的 $L_j = 2$ 时, (3.114)与(3.115)式分别简化为(3.112)与(3.113)式.

上述方法在文献中常称为“折层”法 (the method of “collapsed strata”).

3.7.4 事后分层

在分层抽样中, 一般的必须在抽样前就将总体中的全部抽样单元分

好层. 如果事先分层有困难, 譬如说缺少总体单元按层的抽样框或因 N 太大事先分层太费事, 或调查前每个单元属于哪一层不清楚等等情形, 若要利用分层抽样的优点, 就应采用对样本的事后分层(poststratification)技术.

事后分层是先用简单随机抽样从总体中抽取一个样本量为 n 的样本, 然后再对样本中的单元按某些特征进行分层. 若记属于第 h 层的单元数为 $n_h \left(\sum_{h=1}^L n_h = n \right)$, 则只要 $W_h = N_h/N$ 可通过其他途径得到, 则对总体均值 \bar{Y} 的事后分层可估计为:

$$\bar{y}_{\text{post}} = \sum_{h=1}^L W_h \bar{y}_h \quad (3.116)$$

其中

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}. \quad (3.117)$$

为求 \bar{y}_{post} 的方差, 首先注意到在 n_h 固定且都大于 0 的条件下, $\{y_{h1}, \dots, y_{hn_h}\}$ ($h=1, 2, \dots, L$) 可看成是独立的从各层中抽取的简单随机样本. 事实上, 不妨设 $L=2$, 以事件 A 表示“在第 1 层中抽到 y_{11}, \dots, y_{1n_1} ”, 事件 B 表示“在第 2 层中抽到 y_{21}, \dots, y_{2n_2} ”; B' 表示“在第 2 层中抽到 n_2 单元”, 则

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{1 / \binom{N}{n}}{\binom{N_1}{n_1} / \binom{N}{n}} = \frac{1}{\binom{N_1}{n_1}}.$$

同理可证

$$P(A|B') = \frac{1}{\binom{N_1}{n_1}}.$$

故有

$$P(A|B) = P(A|B') = \frac{1}{\binom{N_1}{n_1}}.$$

这表明在第 1 层中抽到 y_{11}, \dots, y_{1n_1} 的概率与在第 2 层中哪 n_2 单元被抽到无关. 同时也说明 $\{y_{11}, \dots, y_{1n_1}\}$ 可看成是从第 1 层抽取的简单随机样本. 因而在 n_h 固定且皆大于 0 的条件下, 有

$$V(\bar{y}_{\text{post}}) = \sum_h \frac{W_h^2 S_h^2}{n_h} - \frac{1}{N} \sum_h W_h S_h^2, \quad (3.118)$$

其中

$$S_A^2 = \frac{1}{N_A - 1} \sum_{i=1}^{N_A} (Y_{Ai} - \bar{Y}_A)^2. \quad (3.119)$$

定理 3.9 当 n 充分大时, 事后分层估计量 \bar{y}_{pst} 及其方差 $V(\bar{y}_{\text{pst}})$ 有性质:

$$1) E(\bar{y}_{\text{pst}}) = \bar{Y}; \quad (3.120)$$

$$2) E[V(\bar{y}_{\text{pst}})] \approx \frac{1-f}{n} \sum_h W_h S_h^2 + \frac{1}{n^2} \sum_h (1-W_h) S_h^2. \quad (3.121)$$

证明 当 n 充分大时, 可以认为 $n_h > 0$ ($h=1, \dots, L$), 注意到当 n_h 固定时, y_h 是 \bar{Y}_h 的无偏估计, 因而

$$\begin{aligned} E(\bar{y}_{\text{pst}}) &= \sum_h W_h E(y_h) = \sum_h W_h E[E(\bar{y}_h | n_h \text{ 固定})] \\ &= E\left(\sum_h W_h \bar{Y}_h\right) = \bar{Y}. \end{aligned}$$

Setphan(1945)证明了以下的结果.

$$E\left[\frac{1}{n_h}\right] = \frac{1}{nW_h} + \frac{1}{n^2 W_h^2} + O\left(\frac{1}{n^4}\right). \quad (3.122)$$

因此对大的 n , 有

$$E\left[\frac{1}{n_h}\right] \approx \frac{1}{nW_h} + \frac{1}{n^2 W_h^2}.$$

从而

$$\begin{aligned} E[V(\bar{y}_{\text{pst}})] &\approx \frac{1}{n} \sum_h W_h S_h^2 + \frac{1}{n^2} \sum_h (1-W_h) S_h^2 - \frac{1}{N} \sum_h W_h S_h^2 \\ &= \frac{1-f}{n} \sum_h W_h S_h^2 + \frac{1}{n^2} \sum_h (1-W_h) S_h^2 \\ &= V_{\text{prop}} + \frac{1}{n^2} \sum_h (1-W_h) S_h^2. \end{aligned} \quad (3.123)$$

其中第一项恰是比例分配分层抽样估计量的方差, 而第二项则表示因事后分层引起的方差的增加量. 由此可见, 当 n 足够大时, 事后分层的精度相当于比例分配事先分层的精度.

事后分层技术有重要的实际意义. 在许多实际问题中经常需要按不同分类的统计数字. 若采用事先的多重分层方法, 困难较大, 而且实际上也不需要每个子层的估计. 于是可对一个从总体中抽取的简单随机样本按每一种分层方式进行事后分层, 只要按每种分层的总体数(或层权)已知, 即可获得按这种分类的事后分层估计量. 另一方面, 从原则上说, 事后分层也可用于按某一种(或两种)事先分层, 但严格比例分配的样本. 因为这种样本与简单随机样本一样, 是自加权的, 总体中每个单元被抽中的

概率都相等。

3.7.5 定额抽样

在社会调查诸如民意测验、市场调查中,有时采用这样一种快速调查方法.将调查对象按性别、年龄段、职业、受教育程度等分类,事先按比例分配确定每一组对象需要调查的样本量.实际调查时并不使用抽样框进行严格的随机抽样,而是在一定范围内抽样,将抽到的每个对象纳入适当的层中,直到每层都达到所需的样本量 n_h 为止,这就是所谓的定额抽样 (quota sampling)。

由于各层样本量都已事先固定,所以由上一段中关于各层子样本的独立性与随机性的说明可知,如果在具体抽样时是完全随机的(在整个总体范围内),则定额抽样实质上相当于分层随机抽样.不过在实际抽样中,往往只是在一个较小范围内进行,而不是在总体范围内进行.因此,定额抽样在每个层内的抽样或多或少地带有某种非随机性.所以通常的定额抽样并不是一种严格的概率抽样,因而常遭到非议.关于定额抽样与概率抽样的比较,有兴趣的读者可参考 Stephan 与 McCarthy(1958)的文章。

定额抽样在实施过程中的实际工作量比一般想象的要大,因为越往后,抽到“无用”样本的可能性越大.也即当抽样到后面阶段时,大部分被抽到的单元可能是属于早已满额的那些层的,而这些单元只能弃之不用.这反过来会促使调查者有意去挑选指定层的单元,而这又破坏了随机性的原则。

因为“定额”通常是按比例分配的,因此,定额抽样所得的样本可以看成是自加权的.从而它的数据处理非常简单,这也是它在简单而快速的调查中,受到欢迎而乐于被采用的主要原因。

§ 3.8 用于分层的二相抽样

3.8.1 层权误差对分层估计的影响

迄今为止,我们都假定层权 W_h 是已知的.如果 W_h 未知而又不能较精确地估计时,将对分层估计量带来严重的影响。

设使用的层权是 W'_h , 采用以下的估计:

$$\hat{\bar{Y}} = \sum_h W'_h y_h.$$

此时 $\hat{\bar{Y}}$ 不再是无偏的, 其偏倚为 $\sum_h (W'_h - W_h) \bar{Y}_h$, 而且它不因 n 的增大而减小. 这就是说, 此时 $\hat{\bar{Y}}$ 不再是一个可用的估计量. $\hat{\bar{Y}}$ 的均方误差为:

$$\text{MSE}(\hat{\bar{Y}}) = \sum_h \frac{W_h'^2 S_h^2}{n_h} (1 - f_h) + [\sum_h (W'_h - W_h) \bar{Y}_h]^2.$$

因此当 W_h 有误差时, 因分层在精度上的得益将随着 n 的增大而迅速丧失. 当 n 超过一定量时, 分层估计的均方误差就可能比简单随机抽样的简单估计的方差还要大.

例 3.9 考虑一个 $S^2 = 1$ 的简单总体, 分为两层, $W_1 = 0.9$, $W_2 = 0.1$. 假定 $S_1^2 = S_2^2 = S_w^2$, 又 N_h 都很大. 根据 (3.41) 式, 有

$$\begin{aligned} S^2 &= \sum_h W_h S_h^2 + \sum_h W_h (\bar{Y}_h - \bar{Y})^2 \\ &= S_w^2 + W_1 (\bar{Y}_1 - \bar{Y})^2 + W_2 (\bar{Y}_2 - \bar{Y})^2 \\ &= S_w^2 + W_1 (\bar{Y}_1 - W_1 \bar{Y}_1 - W_2 \bar{Y}_2)^2 + W_2 (\bar{Y}_2 - W_1 \bar{Y}_1 - W_2 \bar{Y}_2)^2 \\ &= S_w^2 + W_1 W_2 (\bar{Y}_1 - \bar{Y}_2)^2, \end{aligned}$$

即
$$1 = S_w^2 + 0.09 (\bar{Y}_1 - \bar{Y}_2)^2.$$

若令 $\bar{Y}_1 - \bar{Y}_2 = 1$, 则 $S_w^2 = 0.91$, 因此, 当层权正确时, 按比例分配的分层抽样估计量的方差比简单随机抽样减少 0.09; 若令 $\bar{Y}_1 - \bar{Y}_2 = 3$, 则 $S_w^2 = 0.19$, 此时分层抽样的方差比简单随机抽样的方差减少 0.81. 后者因分层而在精度上的得益比前者为高.

现在考虑用不正确的层权 W'_1 与 W'_2 , 偏倚可以写成:

$$\begin{aligned} \sum_h (W'_h - W_h) Y_h &= (W'_1 - W_1) \bar{Y}_1 + (W'_2 - W_2) \bar{Y}_2 \\ &= (W'_1 - W_1) \bar{Y}_1 - (W'_1 - W_1) \bar{Y}_2 \\ &= (W'_1 - W_1) (\bar{Y}_1 - \bar{Y}_2). \end{aligned}$$

设 $W'_1 = 0.92$, $W'_2 = 0.08$, 则上述两种分层抽样与简单随机抽样可作如下的比较:

分层随机抽样 I:

$$B = 0.02, \quad \text{MSE} = \frac{0.91}{n} + 0.0004;$$

分层随机抽样 II:

$$B = 0.06, \quad \text{MSE} = \frac{0.19}{n} + 0.0036;$$

简单随机抽样:

$$B = 0, \text{MSE} = \frac{1}{n}.$$

对不同的 n , 均方误差 MSE 的值见表 3.16. 从表中可看出, 对不正确的层权, n 愈大, 精度损失愈大, 且对高效的分层, 损失更大.

表 3.16 层权不正确时分层随机抽样均方误差的比较

n	简单随机抽样	分层抽样 I (低效)	分层抽样 II 高效
50	0.0200	0.0188	0.0074
200	0.0050	0.0049	0.0045
400	0.0025	0.0027	0.0041

3.8.2 二相抽样及估计量均值与方差的一般公式

如果层权未知, 如何正确应用分层技术呢? 此时一个可以替代的办法是先从总体中抽取一个相对比较大的简单随机样本, 对这个样本并不需要测定样本单元的指标值, 而仅是将单元按分层特性进行分类, 也即判定单元所属的层. 因此抽取这个第一相样本(the first phase sample)的目的仅是为了估计层权 W_h . 然后在第一相样本中按分层抽样抽取一个相对比较小的子样本——第二相样本(the second phase sample). 对这个子样本作实际调查, 测定其中每个单元的指标值, 再按一般分层抽样的方法作出总体目标量的估计. 这就是用于分层的二相抽样(two-phase sampling), 也称二重抽样(double sampling).

二相抽样也可以用于其他目的. 二相抽样中的第一相样本都是为了获得估计所需的有关总体的辅助信息, 而第二相样本是从第一相样本中抽取的子样本, 才是用来作实际调查的. 作二相抽样的目的是为了提高估计的精度. 显然, 因为需要抽取第一相样本而耗费部分费用, 因此第一相样本的抽取及处理(例如对样本单元进行分类), 相对地说必须是廉价的. 由于从第一相样本中获得的信息而使估计量在精度上的改善必须超过由于已耗费部分费用而不得不减少第二相样本的样本量所造成的精度上的损失. 这是采用二相抽样的必要前提.

二相抽样是一种二步抽样或二次抽样. 为了讨论二相抽样估计量的均值与方差性质, 在这里我们给出在一般的二步抽样中, 估计量 $\hat{\theta}$ 的均值、方差(或协方差)的一般表达式. 此时均值与方差也必须分两步进行. 首先是在给定第一个样本的条件下对第二步抽样求均值和方差, 分

别记为 E_2 与 V_2 ; 然后再对第一步抽样求均值和方差, 记为 E_1 与 V_1 , 因此对于 $\hat{\theta}$ 的均值, 有

$$E(\hat{\theta}) = E_1[E_2(\hat{\theta})]. \quad (3.124)$$

对于 $\hat{\theta}$ 的方差或 $\hat{\theta}_1, \hat{\theta}_2$ 的协方差, 我们有以下引理:

引理 3.1 对任何一个二步抽样, 下列两式成立:

$$1) \quad V(\hat{\theta}) = V_1[E_2(\hat{\theta})] + E_1[V_2(\hat{\theta})]; \quad (3.125)$$

$$2) \quad \text{Cov}(\hat{\theta}_1, \hat{\theta}_2) = \text{Cov}_1[E_2(\hat{\theta}_1), E_2(\hat{\theta}_2)] + E_1[\text{Cov}_2(\hat{\theta}_1, \hat{\theta}_2)]. \quad (3.126)$$

证明 令 $\tilde{\theta} = E(\hat{\theta})$, 则

$$V(\hat{\theta}) = E(\hat{\theta} - \tilde{\theta})^2 = E_1[E_2(\hat{\theta} - \tilde{\theta})^2].$$

$$\begin{aligned} \text{因为} \quad E_2(\hat{\theta} - \tilde{\theta})^2 &= E_2(\hat{\theta}^2) - 2\tilde{\theta}E_2(\hat{\theta}) + \tilde{\theta}^2 \\ &= [E_2(\hat{\theta})]^2 + V_2(\hat{\theta}) - 2\tilde{\theta}E_2(\hat{\theta}) + \tilde{\theta}^2, \end{aligned}$$

两边求 E_1 , 注意到 $\tilde{\theta} = E_1[E_2(\hat{\theta})]$, 即有

$$\begin{aligned} V(\hat{\theta}) &= E_1[E_2(\hat{\theta})]^2 + E_1[V_2(\hat{\theta})] - [E_1E_2(\hat{\theta})]^2 \\ &= V_1[E_2(\hat{\theta})] + E_1[V_2(\hat{\theta})]. \end{aligned}$$

从而(3.125)式成立. (3.126)式的证明与此完全类似, 留给读者作为练习.

引理 3.1 也可推广到多步抽样, 例如对于三步抽样, 有

$$V(\hat{\theta}) = V_1E_2E_3(\hat{\theta}) + E_1V_2E_3(\hat{\theta}) + E_1E_2V_3(\hat{\theta}). \quad (3.127)$$

引理 3.2 设 \bar{y}' 是从总体中抽取的样本量为 n' 的简单随机样本的均值, \bar{y}_1 是从上述样本中抽取的样本量为 n_1 的简单随机子样本 $\{y_1, \dots, y_{n_1}\}$ 的均值, \bar{y}_2 是在第一个样本中剩下的 $n_2 = n' - n_1$ 个单元 (不妨记为 $y_{n_1+1}, \dots, y_{n'}$) 所组成的子样本的均值, 则

$$1) \quad V(\bar{y}_1 - \bar{y}') = S^2 \left(\frac{1}{n_1} - \frac{1}{n'} \right); \quad (3.128)$$

$$2) \quad \text{Cov}(\bar{y}', \bar{y}_1 - \bar{y}') = 0. \quad (3.129)$$

其中 S^2 是总体方差.

证明 首先我们注意两个基本事实: 一是由于 $\{y_1, \dots, y_{n_1}\}$ 是从第一个样本中抽取的简单随机子样本, 那么剩下的单元 $\{y_{n_1+1}, \dots, y_{n'}\}$ 也可以看成是它的一个样本量为 $n' - n_1 = n_2$ 的简单随机子样本 (当然与第一个子样本不是独立的); 其次由于两步抽样都是简单随机的, 因此这两个简单随机子样本可以看作是从总体中直接抽取的简单随机样本. 于是

$$V(\bar{y}_1) = \frac{N-n_1}{n_1 N} S^2, \quad V(\bar{y}_2) = \frac{N-n_2}{n_2 N} S^2.$$

上述两个公式也可根据引理 3.1 得到, 因为

$$E_2(\bar{y}_1) = \bar{y}', \quad V_2(\bar{y}_1) = \frac{n' - n_1}{n_1 n'} s'^2,$$

其中

$$s'^2 = \frac{1}{n' - 1} \sum_{i=1}^{n'} (y_i - \bar{y}')^2.$$

所以从引理 3.1 得

$$\begin{aligned} V(\bar{y}_1) &= V_1 E_2(\bar{y}_1) + E_1 V_2(\bar{y}_1) \\ &= V_1(\bar{y}') + \frac{n' - n_1}{n_1 n'} E_1(s'^2) \\ &= \frac{N - n'}{n' N} S^2 + \frac{n' - n_1}{n_1 n'} S^2 = \frac{N - n_1}{n_1 N} S^2. \end{aligned}$$

注意到

$$V(\bar{y}') = \frac{N - n'}{n' N} S^2,$$

另一方面, 又有

$$\begin{aligned} V(\bar{y}') &= V\left(\frac{n_1}{n'} \bar{y}_1 + \frac{n_2}{n'} \bar{y}_2\right) \\ &= \frac{n_1^2}{n'^2} V(\bar{y}_1) + \frac{n_2^2}{n'^2} V(\bar{y}_2) + \frac{2n_1 n_2}{n'^2} \text{Cov}(\bar{y}_1, \bar{y}_2), \end{aligned}$$

故

$$\text{Cov}(\bar{y}_1, \bar{y}_2) = -\frac{S^2}{N}.$$

因为

$$\bar{y}_1 - \bar{y}' = \frac{n_2}{n'} (\bar{y}_1 - \bar{y}_2),$$

故

$$\begin{aligned} V(\bar{y}_1 - \bar{y}') &= \frac{n_2^2}{n'^2} [V(\bar{y}_1) + V(\bar{y}_2) - 2 \text{Cov}(\bar{y}_1, \bar{y}_2)] \\ &= \frac{n_2^2}{n'^2} S^2 \left(\frac{N - n_1}{n_1 N} + \frac{N - n_2}{n_2 N} + \frac{2}{N} \right) \\ &= \left(\frac{1}{n_1} - \frac{1}{n'} \right) S^2. \end{aligned}$$

$$\text{Cov}(\bar{y}', \bar{y}_1 - \bar{y}')$$

$$= \text{Cov} \left[\frac{1}{n'} (n_1 \bar{y}_1 + n_2 \bar{y}_2), \frac{n_2}{n'} (\bar{y}_1 - \bar{y}_2) \right]$$

$$= \frac{1}{n'^2} [n_1 n_2 V(\bar{y}_1) - n_1 n_2 \text{Cov}(\bar{y}_1, \bar{y}_2)$$

$$+ n_2^2 \text{Cov}(\bar{y}_1, \bar{y}_2) - n_2^2 V(\bar{y}_2)] = 0. \quad \blacksquare$$

3.8.3 用于分层的二相抽样的估计

用于分层的二相抽样最早是由 Neyman(1938)提出的. 若总体单元分层的原理是明确的, 但层权 W_h 未知, 则可从总体中先抽取一个样本量为 n' 的简单随机样本进行估计. 令 n'_h 是该第一相样本中属于 h 层的单元数, 则

$$w_h = \frac{n'_h}{n'} \quad (3.130)$$

可作为 W_h 的估计. 如果将总体单元属于 h 层看作是单元的一种特征, 那么 w_h 与 W_h 即是样本及总体中具有这种特征单元的比例. 因而根据第二章的讨论知, w_h 是 W_h 的无偏估计.

第二相抽样是在第一相样本中进行分层随机抽样, 每层中的抽样比 $\nu_h = \frac{n_h}{n'_h}$ 事先指定. 记 $n = \sum_{h=1}^L n_h$ 为第二相样本量, $\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ 是 h 层中第二相样本的均值, 则二相抽样对总体均值 \bar{Y} 的估计为

$$\bar{y}_{st} = \sum_{h=1}^L w_h \bar{y}_h. \quad (3.131)$$

定理 3.10 对于二相分层样本, 若 n' 足够大, 使每个 $w_h > 0$, 又二相样本抽样比 $\nu_h = n_h/n'_h$ 皆事先指定, 则

$$1) \quad E(\bar{y}_{st}) = \bar{Y}; \quad (3.132)$$

$$2) \quad V(\bar{y}_{st}) = S^2 \left(\frac{1}{n'} - \frac{1}{N} \right) + \sum_h \frac{W_h S_h^2}{n'} \left(\frac{1}{\nu_h} - 1 \right). \quad (3.133)$$

证明 记 y'_{hi} 是第一相样本 h 层中的指标值, $\bar{y}'_h = \frac{1}{n'_h} \sum_{i=1}^{n'_h} y'_{hi}$ 是它的均值. 由于事实上并未对所有的第一相样本单元进行 y 的测量, 因此 \bar{y}'_h 是未知的. 令

$$\bar{y}' = \sum_h w_h \bar{y}'_h,$$

则它实际上是第一相样本均值. 对 \bar{y}_{st} 用二步求均值法:

$$\begin{aligned} E(\bar{y}_{st}) &= E_1 E_2(\bar{y}_{st}) = E_1 E_2 \left(\sum_h w_h \bar{y}_h \right) \\ &= E_1 \left[\sum_h w_h E_2(\bar{y}_h) \right] = E_1 \left[\sum_h w_h \bar{y}'_h \right] \\ &= E_1(\bar{y}') = \bar{Y}. \end{aligned}$$

从而 \bar{y}_{st} 是无偏的. 为求 \bar{y}_{st} 的方差, 将它改写成

$$\bar{y}_{st} = \sum_h w_h \bar{y}_h = \sum_h w_h \bar{y}'_h + \sum_h w_h (\bar{y}_h - \bar{y}'_h) = \bar{y}' + \sum_h w_h (\bar{y}_h - \bar{y}'_h).$$

根据引理 3.1 得

$$\begin{aligned}
 V(\bar{y}_{st}) &= V(\bar{y}') + V\left[\sum_h w_h(\bar{y}_h - \bar{y}_h')\right] + 2 \operatorname{Cov}\left[\bar{y}', \sum_h w_h(\bar{y}_h - \bar{y}_h')\right] \\
 &= V(\bar{y}') + V_1 E_2\left[\sum_h w_h(y_h - \bar{y}_h')\right] + E_1 V_2\left[\sum_h w_h(\bar{y}_h - \bar{y}_h')\right] \\
 &\quad + 2 \operatorname{Cov}_1[\bar{y}', E_2(\sum_h w_h(\bar{y}_h - \bar{y}_h'))] \\
 &\quad + 2 E_1[\operatorname{Cov}_2(\bar{y}', \sum_h w_h(\bar{y}_h - \bar{y}_h'))] \\
 &= V(\bar{y}') + E_1 V_2\left[\sum_h w_h(y_h - \bar{y}_h')\right]. \quad (3.134)
 \end{aligned}$$

最后一个等式成立, 是因为在第一相样本固定情况下, \bar{y}' 是常数, 因而

$$\operatorname{Cov}_2[\bar{y}', \sum_h w_h(y_h - \bar{y}_h')] = 0,$$

同时 $E_2\left[\sum_h w_h(\bar{y}_h - \bar{y}_h')\right] = 0$. 另外

$$\begin{aligned}
 &V_2\left[\sum_h w_h(y_h - \bar{y}_h')\right] \\
 &= \sum_h w_h^2 V_2(y_h) = \sum_h w_h^2 \left(\frac{1}{n_h} - \frac{1}{n_h'}\right) s_h'^2 \\
 &= \sum_h \frac{w_h}{n_h'} \left(\frac{1}{\nu_h} - 1\right) s_h'^2.
 \end{aligned}$$

这里 $s_h'^2 = \frac{1}{n_h' - 1} \sum_{i=1}^{n_h'} (y_{hi} - \bar{y}_h')^2$, 因而

$$\begin{aligned}
 &E_1 V_2\left[\sum_h w_h(\bar{y}_h - \bar{y}_h')\right] \\
 &= \frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1\right) E_1(w_h s_h'^2) \\
 &= \frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1\right) E_1 E(w_h s_h'^2 \mid w_h \text{ 固定}) \\
 &= \frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1\right) E(w_h S_h^2) \\
 &= \frac{1}{n'} \sum_h \left(\frac{1}{\nu_h} - 1\right) W_h S_h^2.
 \end{aligned}$$

而
$$V(\bar{y}') = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2.$$

故从(3.134)式得到

$$V(\bar{y}_{st}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \sum_h \frac{W_h S_h^2}{n'} \left(\frac{1}{\nu_h} - 1\right). \quad \blacksquare$$

$V(\bar{y}_{st})$ 有多种表达形式, 根据(3.41)式, 有

$$(N-1)S^2 = \sum_h (N_h-1)S_h^2 + \sum_h N_h(\bar{Y}_h - \bar{Y})^2.$$

记 $g' \triangleq (N - n') / (N - 1)$, 则若在上式两边同时乘以 $g' / (n'N)$, 可得

$$\left(\frac{1}{n'} - \frac{1}{N}\right) S^2 = \frac{g'}{n'} \sum_h \left(W_h \cdot \frac{1}{N}\right) S_h^2 + \frac{g'}{n'} \sum_h W_h (\bar{Y}_h - Y)^2,$$

将此式代入(3.133)式, 得

$$\begin{aligned} V(\bar{y}_{st}) &= \sum_h W_h S_h^2 \left(\frac{1}{n'v_h} - \frac{1}{N}\right) + \frac{g'}{n'N} \sum_h (W_h - 1) S_h^2 \\ &\quad + \frac{g'}{n'} \sum_h W_h (\bar{Y}_h - \bar{Y})^2. \end{aligned} \quad (3.135)$$

通常, (3.135)式中的第二项数值较小, 可以忽略不计, 因而有以下的近似公式:

$$V(\bar{y}_{st}) \approx \sum_h W_h S_h^2 \left(\frac{1}{n'v_h} - \frac{1}{N}\right) + \frac{g'}{n'} \sum_h W_h (\bar{Y}_h - \bar{Y})^2. \quad (3.136)$$

下面讨论 $V(\bar{y}_{st})$ 的估计问题. 当 n' 与 N 都很大, 从而 $1/N$ 与 $1/n'$ 都很小时, 根据(3.136)式, 可得它的一个几乎是无偏的估计量:

$$v(\bar{y}_{st}) \approx \sum_h w_h s_h^2 \left(\frac{1}{n'v_h} - \frac{1}{N}\right) + \frac{g'}{n'} \sum_h w_h (\bar{y}_h - \bar{y}_{st})^2. \quad (3.137)$$

$$\approx \sum_h w_h^2 s_h^2 \left(\frac{1}{n_h} - \frac{1}{n_h'}\right) + \left(\frac{1}{n'} - \frac{1}{N}\right) \sum_h w_h (\bar{y}_h - \bar{y}_{st})^2. \quad (3.138)$$

在几乎所有的应用场合, 上述的公式都是够用的. 然而当 $1/n'$ 与 $1/N$ 都不可忽略时, 结果就比较复杂. 此时我们有如下的定理:

定理 3.11 对于分层二相抽样, $V(\bar{y}_{st})$ 的一个无偏估计量为:

$$\begin{aligned} v(\bar{y}_{st}) &= \frac{n'(N-1)}{(n'-1) \cdot N} \left[\sum_h w_h s_h^2 \left(\frac{1}{n'v_h} - \frac{1}{N}\right) \right. \\ &\quad \left. + \frac{g'}{n'} \sum_h s_h^2 \left(\frac{w_h}{N} - \frac{1}{n'v_h}\right) \right. \\ &\quad \left. + \frac{g'}{n'} \sum_h w_h (\bar{y}_h - \bar{y}_{st})^2 \right]. \end{aligned} \quad (3.139)$$

证明

$$\begin{aligned} E \left[\sum_h w_h s_h^2 \left(\frac{1}{n'v_h} - \frac{1}{N}\right) \right] &= \sum_h \left(\frac{1}{n'v_h} - \frac{1}{N}\right) E[E(w_h s_h^2 | w_h \text{ 固定})] \\ &= \sum_h \left(\frac{1}{n'v_h} - \frac{1}{N}\right) E(w_h S_h^2) \\ &= \sum_h \left(\frac{1}{n'v_h} - \frac{1}{N}\right) W_h S_h^2. \end{aligned} \quad (3.140)$$

同理可证

$$E \left[\frac{g'}{n'} \sum_h s_h^2 \left(\frac{w_h}{N} - \frac{1}{n'v_h}\right) \right] = \frac{g'}{n'} \sum_h S_h^2 \left(\frac{W_h}{N} - \frac{1}{n'v_h}\right). \quad (3.141)$$

为求

$$\sum_h w_h (y_h - \bar{y}_{st})^2 = \sum_h w_h \bar{y}_h^2 - \bar{y}_{st}^2$$

的均值, 在固定 w_h 的情形下, 有

$$\begin{aligned} E(\sum_h w_h \bar{y}_h^2) &= \sum_h w_h E(\bar{y}_h^2) = \sum_h w_h [E(\bar{y}_h)^2 + V(\bar{y}_h)] \\ &= \sum_h w_h [\bar{Y}_h^2 + V(\bar{y}_h - \bar{y}'_h) + V(\bar{y}'_h) + 2\text{Cov}(\bar{y}'_h, \bar{y}_h - \bar{y}'_h)] \\ &= \sum_h w_h \left[\bar{Y}_h^2 + S_h^2 \left(\frac{1}{n_h} - \frac{1}{n'_h} \right) + S_h^2 \left(\frac{1}{n'_h} - \frac{1}{N_h} \right) \right] \\ &= \sum_h w_h \left[\bar{Y}_h^2 + S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \right] \\ &= \sum_h w_h \bar{Y}_h^2 + \sum_h S_h^2 \left(\frac{1}{n'_h n'} - \frac{w_h}{N_h} \right). \end{aligned} \quad (3.142)$$

再对 w_h 求均值, 得到

$$E(\sum_h w_h \bar{y}_h^2) = \sum_h W_h \bar{Y}_h^2 + \sum_h S_h^2 \left(\frac{1}{n'_h n'} - \frac{1}{N} \right), \quad (3.143)$$

而
故

$$E(\bar{y}_{st}^2) = Y^2 + V(\bar{y}_{st}),$$

$$\begin{aligned} E[\sum_h w_h (\bar{y}_h - \bar{y}_{st})^2] &= \sum_h W_h \bar{Y}_h^2 + \sum_h S_h^2 \left(\frac{1}{n'_h n'} - \frac{1}{N} \right) - \bar{Y}^2 - V(\bar{y}_{st}) \\ &= \sum_h W_h (\bar{Y}_h - \bar{Y})^2 + \sum_h S_h^2 \left(\frac{1}{n'_h n'} - \frac{1}{N} \right) - V(\bar{y}_{st}). \end{aligned} \quad (3.144)$$

于是, 综合(3.140)、(3.141)与(3.144)三式, 即有

$$\begin{aligned} E[v(\bar{y}_{st})] &= \frac{n'(N-1)}{(n'-1)N} \left[\sum_h \left(\frac{1}{n'_h n'} - \frac{1}{N} \right) W_h S_h^2 \right. \\ &\quad + \frac{g'}{n'} \sum_h S_h^2 \left(\frac{W_h}{N} - \frac{1}{n'_h n'} \right) \\ &\quad + \frac{g'}{n'} \sum_h W_h (Y_h - \bar{Y})^2 \\ &\quad \left. + \frac{g'}{n'} \sum_h S_h^2 \left(\frac{1}{n'_h n'} - \frac{1}{N} \right) - \frac{g'}{n'} V(\bar{y}_{st}) \right] \\ &= \frac{n'(N-1)}{(n'-1)N} \left[V(\bar{y}_{st}) - \frac{g'}{n'} V(\bar{y}_{st}) \right] = V(\bar{y}_{st}). \end{aligned}$$

从而 $v(\bar{y}_{st})$ 是 $V(\bar{y}_{st})$ 的无偏估计. ■

第 4 章

比估计与回归估计

前两章中, 对总体参数的估计都是简单(线性)估计, 即对于总体均值的估计, 在简单随机抽样中用的是样本均值(算术平均数); 在分层随机抽样中用的是各层样本均值的加权平均. 在这一章中, 我们研究一些比较复杂的非线性估计, 主要是比估计与回归估计. 此时, 除了调查指标 Y 外, 还有另外的指标, 例如 X 可以利用, X 称为辅助变量(auxiliary variable). 我们利用每个单元的指标值 Y_i 与 X_i 之间的比例关系, 或相关关系来提高对目标量估计的精度. 在实际问题中, X_i 常是 Y_i 的前期资料(例如上一次普查资料), 或对现期 Y_i 的粗略估计, 或表示单元规模的某个量等. 无论何种情形, X 的总体均值 \bar{X} 或总和 X 一般地必须是已知的.

§ 4.1 比估计及其基本性质

4.1.1 定 义

如果 Y_i 与 X_i 之间存在着大致的正比例关系, 则可用比估计量(ratio estimator).

定义 4.1 对于简单随机抽样, 总体均值 \bar{Y} 与总和 Y 的比估计量定义为:

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}} \bar{X} = \frac{y}{x} \bar{X}; \quad (4.1)$$

$$\hat{Y}_R = \frac{y}{x} X = \frac{y}{x} X = N \bar{y}_R. \quad (4.2)$$

其中 \bar{y} , y , \bar{x} , x 分别是样本中 y_i 与 x_i 的平均数与总和.

有时候, 调查的目的就是要估计总体 \bar{Y} 与 \bar{X} 的比值:

$$R = \frac{\bar{Y}}{\bar{X}} = \frac{Y}{X}, \quad (4.3).$$

对它的估计为

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{y}{x}. \quad (4.4)$$

通常比估计是指 y_R 及 \hat{Y}_R , 而 \hat{R} 称为对比值的估计. 由于这三者之间仅相差一个常数, 故在研究它们的性质时, 按照方便, 选用一种进行讨论即可. 许多情况下, 就用 \hat{R} 来说明. 因此本章中凡是涉及比估计的, 都包括 \hat{R} 在内.

4.1.2 基本性质

比估计是有偏的, 但当样本量 n 增大时, 偏倚趋于零. 这就是说, 它是渐近无偏的, 也即当 n 大时, 比估计量可以看成是近似无偏的. 此时均方误差与方差也就近似相等. 事实上, 我们有以下的结果(更精确的结果见 § 4.2).

定理 4.1 对于简单随机抽样, 当 n 大时,

$$E(\bar{y}_R) \approx \bar{Y}, \quad E(\hat{Y}_R) \approx Y, \quad E(\hat{R}) \approx R; \quad (4.5)$$

$$V(\bar{y}_R) \approx \frac{1}{n} f \left[\frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1} \right]; \quad (4.6)$$

$$V(\hat{Y}_R) \approx \frac{N^2(1-f)}{n} \left[\frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1} \right]; \quad (4.7)$$

$$V(\hat{R}) \approx \frac{1}{n\bar{X}^2} f \left[\frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1} \right]. \quad (4.8)$$

证明 只证 \hat{R} 情形: 对于 $\bar{y}_R = \bar{X}\hat{R}$, $\hat{Y}_R = N\bar{X}\hat{R}$, 它们与 \hat{R} 只相差常数因子, 故相应公式可以从有关 \hat{R} 的公式直接推得.

$$\hat{R} - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y}}{\bar{x}} \frac{R\bar{x}}{\bar{x}}, \quad (4.9)$$

当 n 大时, $\bar{x} \approx \bar{X}$, 代入上式分母, 即有

$$E(\hat{R} - R) \approx \frac{1}{\bar{X}} [E(\bar{y}) - RE(\bar{x})] = \frac{1}{\bar{X}} (\bar{Y} - R\bar{X}) = 0.$$

故当 n 大时, $E(\hat{R}) \approx R$. 此时

$$V(\hat{R}) \approx \text{MSE}(\hat{R}) = E(R - \hat{R})^2 \approx \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2.$$

令

$$G_i = Y_i - RX_i \quad (i=1, 2, \dots, N), \quad (4.10)$$

则

$$\begin{aligned} \bar{g} &= \bar{y} - R\bar{x}, \quad \bar{G} = \bar{Y} - R\bar{X} = 0, \\ E(\bar{y} - R\bar{x})^2 &= E(\bar{g}^2) = V(\bar{g}) \end{aligned} \quad (4.11)$$

$$= \frac{1-f}{n} S_g^2 = \frac{1-f}{n} \cdot \frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1}.$$

故
$$V(\hat{R}) \approx \text{MSE}(\hat{R}) \approx \frac{1-f}{n\bar{X}^2} \cdot \frac{\sum_{i=1}^N (Y_i - RX_i)^2}{N-1}.$$

(4.6)~(4.8)式也可用 X_i, Y_i 的方差 S_x^2, S_y^2 和协方差 S_{yx} (或等价地, 用相关系数 ρ) 来表达, 因为

$$\begin{aligned} \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 &= \frac{1}{N-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - R(X_i - \bar{X})]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N [(Y_i - \bar{Y})^2 + R^2(X_i - \bar{X})^2 \\ &\quad - 2R(Y_i - \bar{Y})(X_i - \bar{X})] \\ &= S_y^2 + R^2 S_x^2 - 2RS_{yx} \\ &= S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x \\ &= \bar{Y}^2 (O_y^2 + O_x^2 - 2O_{yx}), \end{aligned} \quad (4.12)$$

其中

$$S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2, \quad S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2, \quad (4.13)$$

$$S_{yx} = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}), \quad (4.14)$$

分别是 Y_i, X_i 的总体方差与协方差, 而

$$\rho = \frac{S_{yx}}{S_y S_x} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2 \cdot \sum_{i=1}^N (X_i - \bar{X})^2}} \quad (4.15)$$

是 Y_i 与 X_i 的相关系数. 又

$$O_y^2 = \frac{S_y^2}{\bar{Y}^2}, \quad O_x^2 = \frac{S_x^2}{\bar{X}^2}, \quad O_{yx} = \frac{S_{yx}}{\bar{Y}\bar{X}} = \rho \frac{S_y S_x}{\bar{Y}\bar{X}} \quad (4.16)$$

分别是 Y_i 及 X_i 的变异系数的平方 (相对方差) 和 Y_i 与 X_i 的相对协方差. 于是当 n 大时,

$$V(\hat{R}) \approx \frac{1-f}{n\bar{X}^2} (S_y^2 + R^2 S_x^2 - 2R\rho S_{yx})$$

$$= \frac{1-f}{n} R^2 (O_y^2 + O_x^2 - 2O_{yx}). \quad (4.17)$$

类似的, 有

$$\begin{aligned} V(\bar{y}_R) &\approx \frac{1}{n} \frac{f}{\bar{Y}^2} (S_y^2 + R^2 S_x^2 - 2R\rho S_{yx}) \\ &\quad - \frac{1}{n} \frac{f}{\bar{Y}^2} \bar{Y}^2 (O_y^2 + O_x^2 - 2O_{yx}); \end{aligned} \quad (4.18)$$

$$\begin{aligned} V(\hat{P}_R) &\approx \frac{N^2(1-f)}{n} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x) \\ &= \frac{1-f}{n} Y^2 (O_y^2 + O_x^2 - 2O_{yx}). \end{aligned} \quad (4.19)$$

鉴于这三个估计量的方差与被估计量的平方只相差同一个常数因子, 故它们的相对方差(变异系数的平方)都相等. 根据前述表达式, 这个量为

$$\begin{aligned} (O_V)^2 &= \frac{V(\hat{R})}{R^2} = \frac{V(\bar{y}_R)}{\bar{Y}^2} = \frac{V(\hat{P}_R)}{Y^2} \\ &\approx \frac{1-f}{n} (O_y^2 + O_x^2 - 2O_{yx}). \end{aligned} \quad (4.20)$$

4.1.3 方差的估计

在前述方差的近似公式中, 都涉及总体的有关量, 因此在实际问题中仍需用样本估计. 我们用

$$\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \quad (4.21)$$

估计 $\frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2$. 这个估计也是有偏的, 但可以证明(详见 §4.2)当 n 大时, 偏倚也趋于 0. 于是 $V(\hat{R})$ 的估计可采用

$$\begin{aligned} v_1(\hat{R}) &= \frac{1-f}{n\bar{X}^2} \cdot \frac{\sum_{i=1}^n (y_i - \hat{R}x_i)^2}{n-1} \\ &= \frac{1}{n\bar{X}^2(n-1)} \left(\sum_{i=1}^n y_i^2 + \hat{R}^2 \sum_{i=1}^n x_i^2 - 2\hat{R} \sum_{i=1}^n y_i x_i \right) \end{aligned} \quad (4.22)$$

$$= \frac{1}{n\bar{X}^2} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx}), \quad (4.23)$$

其中

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2, \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (4.24)$$

$$s_{yx} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}), \quad (4.25)$$

分别是 y_i 与 x_i 的样本方差与协方差.

在 $v_1(\hat{R})$ 中我们也可以用 x_i 的样本均值 \bar{x} 代替总体均值 \bar{X} , 从而得到 $V(\hat{R})$ 的另一种估计:

$$v_2(\hat{R}) = \frac{1-f}{n\bar{x}^2} (s_y^2 + R^2 s_x^2 - 2R s_{yx}). \quad (4.26)$$

注意, $v_2(\hat{R})$ 并不一定不如 $v_1(\hat{R})$, 这两种估计的优劣在不同问题中是不同的. y_R 及 \hat{Y}_R 的方差估计公式可由相应的 $v_1(\hat{R})$ 或 $v_2(\hat{R})$ 得到, 不另行列出了.

4.1.4 置 信 限

对一般的 n , 比估计的分布偏斜程度很大(右偏), 因此在用估计量的正态近似时要特别小心. 通常只有当 $n > 30$, 且 $cv(\bar{x}) < 0.1$, $cv(\bar{y}) < 0.1$ 都得到满足时, 才可直接用正态近似构造置信区间. 否则, 需用以下更为精确的近似:

$$\begin{aligned} E(\bar{y} - R\bar{x}) &= \bar{Y} - R\bar{X} = 0, \\ V(\bar{y} - R\bar{x}) &= \frac{1}{n} f (S_y^2 + R^2 S_x^2 - 2RS_{yx}) \\ &= S_y^2 + R^2 S_x^2 - 2RS_{yx}. \end{aligned}$$

为使符号简洁起见, 这里用 S_y^2 表示 $V(\bar{y})$ 等. 由于只要 n 不太小(例如 $n > 30$), \bar{y}, \bar{x} 即可近似看作遵从正态分布, 从而

$$U = \frac{\bar{y} - R\bar{x}}{\sqrt{S_y^2 + R^2 S_x^2 - 2RS_{yx}}} \quad (4.27)$$

近似遵从 $N(0, 1)$. 对给定置信水平 $1 - \alpha$, 令 $U = \pm u_\alpha$, 即有关于 R 的下列二次方程:

$$\begin{aligned} (\bar{y} - R\bar{x})^2 &= u_\alpha^2 (S_y^2 + R^2 S_x^2 - 2RS_{yx}), \\ (\bar{x}^2 - u_\alpha^2 S_x^2) R^2 - 2(\bar{x}\bar{y} - u_\alpha^2 S_{yx}) R + (\bar{y}^2 - u_\alpha^2 S_y^2) &= 0. \end{aligned}$$

解此二次方程可得 R 的两个根:

$$R = \frac{\hat{R} [(1 - u_\alpha^2 c_{yx}) \pm u_\alpha \sqrt{(c_y^2 + c_x^2 - 2c_{yx}) - u_\alpha^2 (c_y^2 c_x^2 - c_{yx}^2)}]}{(1 - u_\alpha^2 c_x^2)}, \quad (4.28)$$

其中

$$c_y^2 = \frac{s_y^2}{\bar{y}^2} = \frac{1-f}{n} \frac{s_y^2}{\bar{y}^2}, \quad c_x^2 = \frac{s_x^2}{\bar{x}^2} = \frac{1-f}{n} \frac{s_x^2}{\bar{x}^2}, \quad (4.29)$$

$$c_{\bar{y}\bar{x}} = \frac{\frac{s_{\bar{y}\bar{x}}}{\bar{y}\bar{x}}}{\frac{s_{\bar{y}\bar{x}}}{\bar{y}\bar{x}}} = \frac{1-f}{n} \frac{s_{\bar{y}\bar{x}}}{\bar{y}\bar{x}}. \quad (4.30)$$

是样本均值的相对方差和相对协方差。

(4.28)式中的两个 R 值即是 R 在 $1-\alpha$ 置信水平下的置信限。当 $u_{\alpha}^2 c_{\bar{y}}^2, u_{\alpha}^2 c_{\bar{x}}^2, u_{\alpha}^2 c_{\bar{y}\bar{x}}$ 皆很小时 (相应于更大的 n)，(4.28)式即简化为

$$R = \hat{R} \pm u_{\alpha} \hat{R} \sqrt{c_{\bar{y}}^2 + c_{\bar{x}}^2 - 2c_{\bar{y}\bar{x}}}. \quad (4.31)$$

此式即是直接用 \hat{R} 的正态近似所得的 R 的置信限。

4.1.5 比估计与简单估计量的比较

下面的定理给出了在大样本时，比估计比简单估计更为精确的条件。

定理 4.2 对简单随机抽样，若 n 足够大，则当

$$\rho > \frac{1}{2} \frac{S_e/\bar{X}}{S_y/\bar{Y}} = \frac{1}{2} \frac{C_e}{C_y} \quad (4.32)$$

时，有

$$V(\bar{y}_R) < V(\bar{y}).$$

证明 根据定理 4.1，当 n 足够大时

$$V(\bar{y}_R) \approx \frac{1-f}{n} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x),$$

而对简单估计 \bar{y} ，有

$$V(\bar{y}) = \frac{1-f}{n} S_y^2.$$

故当以下关系成立时：

$$R^2 S_x^2 - 2R\rho S_y S_x < 0,$$

即

$$\rho > \frac{RS_x}{2S_y} = \frac{1}{2} \frac{C_e}{C_y},$$

有

$$V(\bar{y}_R) < V(\bar{y}).$$

特别当 $C_e \approx C_y$ (例如当 X_i 是 Y_i 的前期数据时即有此类情况)，只要 $\rho > 1/2$ ，比估计就比简单估计更为精确。

4.1.6 数值例子——小麦估产调查

例 4.1 某县为估计全县的小麦产量，在全县 $N=576$ 个村中用简单随机抽样抽了 $n=24$ 个村。调查了这些村的小麦产量，表 4.1 记录了调查结果及这些村的小麦种植面积。已知全县小麦种植面积总和为

21875.6hm²(公顷).

表 4.1 某县小麦产量调查

村号 i	产量 y_i (t) ^(*)	种植面积 x_i (hm ²)	村号 i	产量 y_i (t)	种植面积 x_i (hm ²)
1	112.0	30.2	13	105.7	30.8
2	129.1	36.1	14	80.5	21.7
3	208.2	60.8	15	163.0	49.2
4	158.5	44.4	16	98.7	28.0
5	110.2	29.8	17	137.8	37.8
6	123.3	34.9	18	141.2	38.6
7	157.7	41.6	19	152.5	42.8
8	154.2	42.8	20	142.5	39.0
9	98.7	25.8	21	136.7	37.6
10	112.7	34.7	22	153.2	43.2
11	125.5	35.1	23	93.0	26.1
12	60.3	15.8	24	179.8	48.3

(*) t 为吨的国际单位符号)

根据表 4.1, 计算得到有关样本的数据如下:

$$n=24, \quad 1-f=0.95833,$$

$$y=\sum_i y_i=3135, \quad x=\sum_i x_i=875.1,$$

$$\bar{y}=\frac{y}{24}=130.625, \quad \bar{x}=\frac{x}{24}=36.4625,$$

$$l_{yy}=\sum_i (y_i-\bar{y})^2=25580.485, \quad s_y^2=\frac{l_{yy}}{23}=1112.195,$$

$$l_{xx}=\sum_i (x_i-\bar{x})^2=2184.65625, \quad s_x^2=\frac{l_{xx}}{23}=94.98505,$$

$$l_{yx}=\sum_i (y_i-\bar{y})(x_i-\bar{x})=7390.1525, \quad s_{yx}=\frac{l_{yx}}{23}=321.3110.$$

于是该县的小麦每公顷平均产量 R 的估计为:

$$\hat{R}=\frac{y}{x}=3.58245(\text{t/hm}^2).$$

小麦总产量 Y 的比估计为:

$$\hat{Y}_R=X\hat{R}=21875.6 \times 3.5824=78368.2(\text{t}).$$

为求 \hat{R} 及 \hat{Y}_R 的方差和标准差的估计, 根据 (4.23) 式, 注意到此时

$$\bar{X}=\frac{21875.6}{576}=37.97847,$$

$$v(\hat{R})=\frac{1-f}{n\bar{X}^2}(s_y^2+\hat{R}^2s_x^2-2\hat{R}s_{yx})$$

$$\begin{aligned}
&= \frac{0.95833}{24 \times 37.97847^2} [1112.195 + 3.58245^2 \times 94.98505 \\
&\quad - 2 \times 3.58245 \times 321.8110] \\
&= 8.0464 \times 10^{-4},
\end{aligned}$$

$$\sqrt{v(\hat{R})} = 0.028366,$$

$$v(\hat{Y}_R) = X^2 v(\hat{R}) = 385054,$$

$$\sqrt{v(\hat{Y}_R)} = 620.5(\text{t}).$$

作为比较,若按简单估计,有

$$\hat{Y} = N\bar{y} = 576 \times 130.625 = 75210(\text{t}),$$

$$v(\hat{Y}) = \frac{N^2(1-f)}{n} s_y^2 = 14734308,$$

$$\sqrt{v(\hat{Y})} = 3838.5(\text{t}).$$

由此可见,在此例中,比估计 \hat{Y}_R 要比简单估计精确多了.

最后我们求 R 的置信水平为 95% 的置信区间,根据(4.28)式,先计

算

$$C_y^2 = \frac{1-f}{n} \frac{s_y^2}{\bar{y}^2} = 2.602746 \times 10^{-3},$$

$$C_x^2 = \frac{1-f}{n} \frac{s_x^2}{\bar{x}^2} = 2.852766 \times 10^{-3},$$

$$C_{yx} = \frac{1-f}{n} \frac{s_{yx}}{\bar{y}\bar{x}} = 2.6937463 \times 10^{-3}.$$

R 的 95% 的上、下置信限为

$$\begin{aligned}
R &= \frac{\hat{R}[(1-u_\alpha^2 c_{yx}^2) \pm u_\alpha \sqrt{(c_y^2 + c_x^2 - 2c_{yx}) - u_\alpha^2(c_y^2 c_x^2 - c_{yx}^2)}]}{1 - u_\alpha^2 c_x^2} \\
&= \frac{3.58245 \times [0.98965 \pm 1.96 \sqrt{(68.0783 - 0.6469) \times 10^{-6}}]}{0.98904} \\
&= 3.6221 \times (0.98965 \pm 0.01609),
\end{aligned}$$

$$R_L = 3.526, R_U = 3.643.$$

即 R 的 95% 的置信区间为 (3.526, 3.643), 相应的 Y 的置信区间为 (77133, 79693).

4.1.7 乘积估计

若辅助变量 \mathcal{X} 与 \mathcal{Y} 呈负相关关系,则不能用比估计,此时应用以下的乘积估计(product estimator);

$$\bar{y}_P = \frac{\bar{x}\bar{y}}{\bar{X}}, \quad \hat{Y}_P = N \frac{\bar{x}\bar{y}}{\bar{X}}. \quad (4.33)$$

与(4.20)式类似, 当 n 大时, 乘积估计的

$$(O_V)^2 \approx \frac{1-f}{n} (O_y^2 + O_x^2 + 2O_{yx}), \quad (4.34)$$

且当

$$\rho < -\frac{1}{2} \frac{S_x/\bar{X}}{S_y/\bar{Y}} = -\frac{1}{2} \frac{O_x}{O_y} \quad (4.33')$$

时, 有

$$V(\bar{y}_P) < V(\bar{y}).$$

§ 4.2 比估计的偏倚及其均方误差和方差估计的阶

4.2.1 关于有限总体样本中心矩阶的基本引理

比估计是有偏的, 它的均方误差或方差也没有无偏的估计量。为了深入研究比估计的偏倚, 均方误差、方差以及它们的一些估计量在大样本时的性质必须考察在 $n \rightarrow \infty$ 时它们趋于零的速度。为此, 在这一小节中我们首先给出在有限总体中抽取的简单随机样本中心矩的基本结果。

若 ξ_n 是 n 的函数, 又

$$\lim_{n \rightarrow \infty} n^k |\xi_n| = K < \infty \quad (K \text{ 为常数}),$$

则记

$$\xi_n = O\left(\frac{1}{n^k}\right).$$

若 $\xi_{n,N}$ 是 n 与 N 的函数

$$\lim_{\substack{n, N \rightarrow \infty \\ n < tN, 0 < t < 1}} n^k |\xi_{n,N}| = K < \infty$$

则

$$\xi_{n,N} = O\left(\frac{1}{n^k}\right).$$

若上面的 $K=0$, 则记 ξ_n 或 $\xi_{n,N} = o\left(\frac{1}{n^k}\right)$ 。

引理 4.1 设 \bar{y} 、 \bar{x} 分别是抽自某有限总体(均值分别为 \bar{Y} 与 \bar{X}) 的简单随机样本的均值, 则对非负整数 k, l , 有

$$E(\bar{y} - \bar{Y})^k = \begin{cases} O(n^{-k/2}), & \text{若 } k \text{ 为偶数;} \\ O(n^{-\frac{k+1}{2}}), & \text{若 } k \text{ 为奇数.} \end{cases} \quad (4.35)$$

$$E[(\bar{y} - \bar{Y})^k (\bar{x} - \bar{X})^l] = \begin{cases} O(n^{-\frac{k+l}{2}}), & \text{若 } k+l \text{ 为偶数;} \\ O(n^{-\frac{k+l+1}{2}}), & \text{若 } k+l \text{ 为奇数.} \end{cases} \quad (4.36)$$

对于绝对中心矩也有相同的结果.

我们不对一般的情况证明上述引理(具体证明可见 David & Sukhatme(1974)的文章), 下面仅指出几种常用且简单的特例:

$$1) \quad E(\bar{y} - \bar{Y})^2 = V(\bar{y}) = O\left(\frac{1}{n}\right).$$

由定理 2.2, 有:

$$E(y - Y)^2 = \frac{N-n}{nN} S_y^2 < \frac{S_y^2}{n} = O\left(\frac{1}{n}\right).$$

$$2) \quad E(\bar{y} - \bar{Y})(\bar{x} - \bar{X}) = \frac{N-n}{nN} S_{yx} = O\left(\frac{1}{n}\right).$$

由定理 2.3 知结论成立.

$$\begin{aligned} 3) \quad E(\bar{y} - \bar{Y})(\bar{x} - \bar{X})^2 &= \frac{(N-n)(N-2n)}{n^2 N(N-1)(N-2)} \\ &\quad \times \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})^2 = O\left(\frac{1}{n^2}\right). \end{aligned} \quad (4.37)$$

特别

$$E(\bar{y} - \bar{Y})^3 = \frac{(N-n)(N-2n)}{n^2 N(N-1)(N-2)} \sum_{i=1}^N (Y_i - \bar{Y})^3 = O\left(\frac{1}{n^2}\right). \quad (4.38)$$

$$\begin{aligned} 4) \quad E(y - \bar{Y})(\bar{x} - \bar{X})^3 &= \frac{N-n}{n^3} \cdot \frac{N^2 + N - 6nN + 6n^2}{N(N-1)(N-2)(N-3)} \\ &\quad \times \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})^3 + \frac{3(n-1)(N-n)(N-n-1)}{n^3(N-1)(N-2)(N-3)} \\ &\quad \times \left[\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}) \right] \left[\sum_{i=1}^N (X_i - \bar{X})^2 \right] = O\left(\frac{1}{n^2}\right). \end{aligned} \quad (4.39)$$

$$\begin{aligned} 5) \quad E(\bar{y} - \bar{Y})^2(\bar{x} - \bar{X})^2 &= \frac{N-n}{n^3} \cdot \frac{N^2 + N - 6nN + 6n^2}{N(N-1)(N-2)(N-3)} \\ &\quad \times \sum_{i=1}^N (Y_i - \bar{Y})^2 (X_i - \bar{X})^2 + \frac{N-n}{n^3} \cdot \frac{(n-1)(N-n-1)}{N(N-1)(N-2)(N-3)} \\ &\quad \times \left\{ \sum_{i=1}^N (Y_i - \bar{Y})^2 \cdot \sum_{i=1}^N (X_i - \bar{X})^2 + 2 \left[\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}) \right]^2 \right\} \\ &= O\left(\frac{1}{n^2}\right). \end{aligned} \quad (4.40)$$

特别

$$E(\bar{y} - \bar{Y})^4 = \frac{N-n}{n^3} \cdot \frac{N^2 + N - 6nN + 6n^2}{N(N-1)(N-2)(N-3)} \sum_{i=1}^N (Y_i - \bar{Y})^4$$

$$+ \frac{3(N-n)(n-1)(N-n-1)}{n^3 N(N-1)(N-2)(N-3)} \left[\sum_{i=1}^N (Y_i - \bar{Y})^2 \right]^2 = O\left(\frac{1}{n^3}\right). \quad (4.41)$$

下面证明(4.37)式. (4.39)及(4.40)式的证明与此类似.

$$\begin{aligned} \text{令} \quad Y_i - \bar{Y} &= U_i, & y_i - \bar{y} &= u_i; \\ X_i - \bar{X} &= V_i, & x_i - \bar{x} &= v_i. \end{aligned} \quad (4.42)$$

$$\begin{aligned} \text{则} \quad U &= 0, & \bar{u} &= \bar{y} - \bar{Y}; \\ \bar{V} &= 0, & \bar{v} &= \bar{x} - \bar{X}. \end{aligned} \quad (4.43)$$

$$\begin{aligned} E(\bar{y} - \bar{Y})(\bar{x} - \bar{X})^2 &= E(\bar{u}v^2) \\ &= \frac{1}{n^3} E\left[\left(\sum_{i=1}^n u_i\right)\left(\sum_{i=1}^n v_i\right)^2\right] \\ &= \frac{1}{n^3} E\left[\left(\sum_{i=1}^n u_i\right)\left(\sum_{i=1}^n v_i^2 + \sum_{i \neq j} v_i v_j\right)\right] \\ &= \frac{1}{n^3} E\left[\sum_{i=1}^n u_i v_i^2 + \sum_{i \neq j} u_i v_j^2 + 2 \sum_{i \neq j} u_i v_i v_j + \sum_{i \neq j \neq k} u_i v_j v_k\right]. \end{aligned} \quad (4.44)$$

根据对称性论证的原理, 有

$$E\left(\sum_{i=1}^n u_i v_i^2\right) = \frac{n}{N} \sum_{i=1}^N U_i V_i^2, \quad (4.45)$$

$$E\left[\sum_{i \neq j} (u_i v_j^2 + 2u_i v_i v_j)\right] = \frac{n(n-1)}{N(N-1)} \sum_{i \neq j} (U_i V_j^2 + 2U_i V_i V_j), \quad (4.46)$$

$$E\left(\sum_{i \neq j \neq k} u_i v_j v_k\right) = \frac{n(n-1)(n-2)}{N(N-1)(N-2)} \sum_{i \neq j \neq k} U_i V_j V_k, \quad (4.47)$$

而

$$\begin{aligned} 0 &= \left(\sum_{i=1}^N U_i\right)\left(\sum_{i=1}^N V_i^2\right) = \sum_{i=1}^N U_i V_i^2 + \sum_{i \neq j} U_i V_j^2 \\ &\Rightarrow \sum_{i \neq j} U_i V_j^2 = -\sum_{i=1}^N U_i V_i^2. \\ 0 &= \left(\sum_{i=1}^N U_i V_i\right)\left(\sum_{i=1}^N V_i\right) = \sum_{i=1}^N U_i V_i^2 + \sum_{i \neq j} U_i V_i V_j \\ &\Rightarrow \sum_{i \neq j} U_i V_i V_j = -\sum_{i=1}^N U_i V_i^2. \\ 0 &= \left(\sum_{i=1}^N U_i\right)\left(\sum_{i=1}^N V_i\right)^2 = \left(\sum_{i=1}^N U_i\right)\left(\sum_{i=1}^N V_i^2 + \sum_{i \neq j} V_i V_j\right) \\ &= \sum_{i=1}^N U_i V_i^2 + \sum_{i \neq j} U_i V_j^2 + 2 \sum_{i \neq j} U_i V_i V_j + \sum_{i \neq j \neq k} U_i V_j V_k \\ &= 2 \sum_{i \neq j} U_i V_i V_j + \sum_{i \neq j \neq k} U_i V_j V_k \end{aligned}$$

$$\Rightarrow \sum_{i \neq j}^N U_i V_j V_k = -2 \sum_{i=1}^N U_i V_i V_i = 2 \sum_{i=1}^N U_i V_i^2.$$

将上述关系代入(4.45)~(4.47)式,从而根据(4.44)式得到:

$$\begin{aligned} E(\bar{y} - \bar{Y})(x - \bar{X})^2 &= \frac{1}{n^2 N} \left[1 - 3 \frac{n-1}{N-1} + 2 \frac{(n-1)(n-2)}{(N-1)(N-2)} \right] \sum_{i=1}^N U_i V_i^2 \\ &= \frac{(N-n)(N-2n)}{n^2 N (N-1)(N-2)} \sum_{i=1}^N U_i V_i^2 \\ &= \frac{(N-n)(N-2n)}{n^2 N (N-1)(N-2)} \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})^2. \end{aligned}$$

4.2.2 比估计的偏倚与均方误差及其阶的估计

定理 4.3 对简单随机抽样, 比估计 $\hat{R} = \frac{\bar{y}}{\bar{x}}$ 的偏倚为:

$$B(\hat{R}) = E(\hat{R} - R) = \frac{1-f}{n} R(C_y^2 - C_x^2) + O\left(\frac{1}{n^2}\right). \quad (4.48)$$

证明 \hat{R} 可写成

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\bar{Y} \left(1 + \frac{\bar{y} - \bar{Y}}{\bar{Y}}\right)}{\bar{X} \left(1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right)} \triangleq R(1 + \delta\bar{y})(1 + \delta\bar{x})^{-1},$$

其中 $\delta\bar{y} = \frac{\bar{y} - \bar{Y}}{\bar{Y}}, \quad \delta\bar{x} = \frac{\bar{x} - \bar{X}}{\bar{X}}, \quad (1 + \delta\bar{x})^{-1} = \frac{\bar{X}}{\bar{x}}.$

于是 \hat{R} 的偏倚可表成:

$$\begin{aligned} B(\hat{R}) &= E(\hat{R} - R) = RE[(\delta\bar{y} - \delta\bar{x})(1 + \delta\bar{x})^{-1}] \\ &= RE\left[(\delta\bar{y} - \delta\bar{x}) \frac{\bar{X}}{\bar{x}}\right]. \end{aligned}$$

而另一方面

$$\begin{aligned} (1 + \delta\bar{x})^{-1} &= 1 - \delta\bar{x} + (\delta\bar{x})^2 - (\delta\bar{x})^3 + \dots \\ &= 1 - \delta\bar{x} + (\delta\bar{x})^2 [1 - (\delta\bar{x}) + (\delta\bar{x})^2 - \dots] \\ &= 1 - \delta\bar{x} + (\delta\bar{x})^2 (1 + \delta\bar{x})^{-1} \\ &= 1 - \delta\bar{x} + (\delta\bar{x})^2 \frac{\bar{X}}{\bar{x}}, \\ B'(\hat{R}) &= RE[(\delta\bar{y} - \delta\bar{x})(1 - \delta\bar{x})] \\ &= RE[\delta\bar{y} - \delta\bar{x} - \delta\bar{y}\delta\bar{x} + (\delta\bar{x})^2] \\ &= R \left[E\left(\frac{(\bar{x} - \bar{X})^2}{\bar{X}^2}\right) - E\left(\frac{(\bar{y} - \bar{Y})(\bar{x} - \bar{X})}{\bar{Y}\bar{X}}\right) \right] \end{aligned}$$

$$\begin{aligned}
 &= R \left[\frac{V(\bar{x})}{\bar{X}^2} - \frac{\text{Cov}(\bar{y}, \bar{x})}{\bar{Y}\bar{X}} \right] \\
 &= \frac{R(1-f)}{n} (C_x^2 - C_{yx}),
 \end{aligned}$$

若令 $X_* = \min_{1 \leq i \leq N} \{X_i\}$, 则根据引理 4.1, 有

$$\begin{aligned}
 |B(\hat{R}) - B'(\hat{R})| &= R \left| E \left[(\delta\bar{y} - \delta\bar{x}) (\delta\bar{x})^2 \frac{\bar{X}}{x} \right] \right| \\
 &\leq \frac{\bar{Y}}{X_*} E \left[\left(\frac{\bar{y}}{\bar{Y}} - \frac{\bar{x} - \bar{X}}{X} \right) \left(\frac{x - \bar{X}}{\bar{X}} \right)^2 \right] = O\left(\frac{1}{n^2}\right).
 \end{aligned}$$

于是
$$B(\hat{R}) = \frac{f}{n} R (C_x^2 - C_{yx}) + O\left(\frac{1}{n^2}\right). \quad \blacksquare$$

注: 由于 $(1 + \delta\bar{x})^{-1} = 1 - \delta\bar{x} + \dots + (\delta\bar{x})^{2k-2} - (\delta\bar{x})^{2k-1} + \dots$,
故若令

$$B_k(\hat{R}) = RE\{(\delta\bar{y} - \delta\bar{x})[1 - \delta\bar{x} + \dots + (\delta\bar{x})^{2k-2} - (\delta\bar{x})^{2k-1}]\},$$

则
$$|B(\hat{R}) - B_k(\hat{R})| = R \left| E \left[(\delta\bar{y} - \delta\bar{x}) (\delta\bar{x})^{2k} \frac{\bar{X}}{x} \right] \right| = O\left(\frac{1}{n^{k+1}}\right).$$

定理 4.4 简单随机抽样比值估计量 \hat{R} 的均方误差为:

$$\text{MSE}(\hat{R}) = \frac{1-f}{n\bar{X}^2} \frac{1}{(N-1)} \sum_{i=1}^N (Y_i - RX_i)^2 + O\left(\frac{1}{n^2}\right). \quad (4.49)$$

证明

$$\begin{aligned}
 \text{由 } (\hat{R} - R)^2 &= \left(\frac{\bar{y} - R\bar{x}}{\bar{X}} \right)^2 = \frac{(\bar{y} - R\bar{x})^2}{\bar{X}^2} \\
 &= \frac{(\bar{y} - R\bar{x})^2}{\bar{X}^2} - \frac{\bar{x}^2 - \bar{X}^2}{\bar{x}^2 \bar{X}^2} (\bar{y} - R\bar{x})^2.
 \end{aligned}$$

从而
$$\text{MSE}(\hat{R}) = E(\hat{R} - R)^2$$

$$\begin{aligned}
 &= \frac{1}{\bar{X}^2} E(\bar{y} - R\bar{x})^2 - E \left[\frac{\bar{x}^2 - \bar{X}^2}{\bar{x}^2 \bar{X}^2} (\bar{y} - R\bar{x})^2 \right] \\
 &= \frac{1}{\bar{X}^2} \cdot \frac{1-f}{n} \frac{1}{(N-1)} \sum_{i=1}^N (Y_i - RX_i)^2 \\
 &\quad - E \left[\frac{\bar{x}^2 - \bar{X}^2}{\bar{x}^2 \bar{X}^2} (\bar{y} - R\bar{x})^2 \right].
 \end{aligned}$$

为估计上式第二项的阶, 记

$$X^* = \max_{1 \leq i \leq N} \{X_i\}, \quad X_* = \min_{1 \leq i \leq N} \{X_i\}.$$

$$\left| E \left[\frac{\bar{x}^2 - \bar{X}^2}{\bar{x}^2 \bar{X}^2} (\bar{y} - R\bar{x})^2 \right] \right| \leq \frac{X^* + \bar{X}}{\bar{X}^2 X_*^2} E[(x - \bar{X})(\bar{y} - R\bar{x})^2]$$

$$= O\left(\frac{1}{n^2}\right).$$

$$\text{从而 } E(\hat{R} - R)^2 = \frac{1-f}{n\bar{X}^2} \frac{1}{(N-1)} \sum_{i=1}^N (Y_i - RX_i)^2 + O\left(\frac{1}{n^2}\right). \quad \blacksquare$$

由(4.48)与(4.49)式, 当 n 大时,

$$\begin{aligned} \frac{E(\hat{R}) - R}{\sqrt{\text{MSE}(\hat{R})}} &\approx \frac{\frac{1-f}{n} R(C_x^2 - C_{yx})}{\sqrt{\frac{1-f}{n} \frac{1}{\bar{X}} \sqrt{\frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2}}} \\ &= \sqrt{\frac{1-f}{n}} \cdot \frac{O_x^2 - C_{yx}}{\sqrt{C_y^2 + C_x^2 - 2C_{yx}}} \\ &= \sqrt{\frac{1-f}{n}} \cdot \frac{S_x}{\bar{X}} \cdot \frac{RS_x - \rho S_y}{\sqrt{S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x}} \\ &= O_x \frac{RS_x - \rho S_y}{\sqrt{S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x}}, \end{aligned} \quad (4.50)$$

其中

$$O_x = \sqrt{\frac{1-f}{n}} \cdot \frac{S_x}{\bar{X}} = \sqrt{\frac{1-f}{n}} O_x \quad (4.51)$$

是 \bar{x} 的变异系数. 由此可知 $\frac{E(\hat{R}) - R}{\sqrt{\text{MSE}(\hat{R})}}$ 的阶为 $O\left(\frac{1}{\sqrt{n}}\right)$, 因而 \hat{R} 是可用估计量. 当 n 足够大时, 偏倚可以忽略, 而 $\text{MSE}(\hat{R}) \approx V(\hat{R})$. L. Kish 等就许多实际问题计算了 $\frac{E(\hat{R}) - R}{\sqrt{\text{MSE}(\hat{R})}}$, 发现在绝大多数问题中, 这个量小于 3%.

Hartley 与 Ross(1954)给出了关于 \hat{R} 偏倚的一个精确公式. 考虑

$$\text{Cov}(\hat{R}, \bar{x}) = E\left(\frac{y}{x} \cdot \bar{x}\right) - E(\hat{R})E(\bar{x}) = Y - \bar{X}E(\hat{R}),$$

$$E(\hat{R}) = \frac{Y}{\bar{X}} - \frac{1}{\bar{X}} \text{Cov}(\hat{R}, \bar{x}) = R - \frac{1}{\bar{X}} \text{Cov}(\hat{R}, \bar{x}).$$

从而

$$E(\hat{R}) - R = -\frac{1}{\bar{X}} \text{Cov}(\hat{R}, \bar{x}). \quad (4.52)$$

例 4.2 对表 4.2 的 $N=6$ 的人为总体, 通过计算所有可能的 $n=4$ 的简单随机样本研究比值估计 \hat{R} 及比估计 \bar{y}_R 的偏倚、均方误差与方差.

表 4.2 一个 $N=6$ 的人为总体

	1	2	3	4	5	6
X_i	2	3	5	5	7	8
Y_i	5	7	10	11	15	18

$$\bar{X} = 5, \quad \bar{Y} = 11, \quad R = 2.2,$$

$$S_x^2 = 5.2, \quad S_y^2 = 23.6, \quad S_{yx} = 11.0.$$

表 4.3 从表 4.2 总体中抽取的所有 $n=4$ 简单随机样本的比估计与回归估计

样本号 j	样本包含的单元号	x	y	\hat{R}	\hat{y}_R	b	\hat{y}_R
1	(1, 2, 3, 4)	3.75	8.25	2.2000	11.0000	1.81480	10.5140
2	(1, 2, 3, 5)	4.25	9.25	2.1765	10.8824	1.94915	10.71186
3	(1, 2, 3, 6)	4.50	10.00	2.2222	11.1111	2.14286	11.07143
4	(1, 2, 4, 5)	4.25	9.50	2.2353	11.1765	2.00000	11.00000
5	(1, 2, 4, 6)	4.50	10.25	2.2778	11.3889	2.16667	11.33333
6	(1, 2, 5, 6)	5.00	11.25	2.2500	11.2500	2.11538	11.25000
7	(1, 3, 4, 5)	4.75	10.25	2.1579	10.7895	1.98040	10.7451
8	(1, 3, 4, 6)	5.00	11.00	2.2000	11.0000	2.16667	11.0000
9	(1, 3, 5, 6)	5.50	12.00	2.1818	10.9091	2.14286	10.9286
10	(1, 4, 5, 6)	5.50	12.25	2.2273	11.1364	2.11905	11.1905
11	(2, 3, 4, 5)	5.00	10.75	2.1500	10.7500	2.00000	10.7500
12	(2, 3, 4, 6)	5.25	11.50	2.1905	10.9524	2.23530	10.94118
13	(2, 3, 5, 6)	5.75	12.50	2.1739	10.8696	2.20340	10.8475
14	(2, 4, 5, 6)	5.75	12.75	2.2174	11.0870	2.15250	11.1356
15	(3, 4, 5, 6)	6.25	13.50	2.1600	10.8000	2.44444	10.4445

表 4.3 列出了全部 $\binom{6}{4} = 15$ 个可能的 $n=4$ 简单随机样本。对每个样本, 计算样本均值 \bar{x} 、 \bar{y} 、比值估计 $\hat{R} = \bar{y}/\bar{x}$ 及 \bar{Y} 的比估计 $\hat{y}_R = \hat{R}\bar{x}$ 。为了以后研究回归估计量的方法, 表中的最后两列还列出了样本回归系数 b 及回归估计 \hat{y}_R (详见 § 4.6)。

$$E(\hat{R}) = \frac{1}{15} \sum_{j=1}^{15} \hat{R}_j = 2.201369,$$

$$B(\hat{R}) = E(\hat{R}) - R = 2.201369 - 2.2 = 0.001369,$$

$$MSE(\hat{R}) = E(\hat{R} - R)^2$$

$$= \frac{1}{15} \left[\sum_{j=1}^{15} \hat{R}_j^2 - 4.4 \sum_{j=1}^{15} \hat{R}_j + (2.2)^2 \times 15 \right]$$

$$= 0.0012575,$$

$$V(\hat{R}) = \text{MSE}(\hat{R}) - B^2(\hat{R}) = 0.0012556,$$

$$\frac{B(\hat{R})}{\sqrt{\text{MSE}(\hat{R})}} = \frac{0.001369}{\sqrt{0.0012575}} = 0.0386.$$

将上面的结果乘以 $\bar{X}=5$ 或 $\bar{X}^2=25$ 即得有关 \bar{y}_R 的偏倚、均方误差和方差, 结果为:

$$B(\bar{y}_R) = 0.006844, \quad \text{MSE}(\bar{y}_R) = 0.03144, \quad V(\bar{y}_R) = 0.03139.$$

4.2.3 均方误差或方差估计的偏倚

在 $\text{MSE}(\hat{R})$ 的主项中包含 $\frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2$, 故在对 $\text{MSE}(\hat{R})$ 或 $V(\hat{R})$ 进行估计时, 自然想到用 $\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2$ 来作为它的估计量. 前已指出, 这个估计量也是有偏的. 事实上, 有以下的定理:

定理 4.5 对简单随机抽样, $\hat{R} = \frac{\bar{y}}{\bar{x}}$ 是 $R = \frac{Y}{X}$ 的估计, 则

$$E\left[\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2\right] = \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 + O\left(\frac{1}{n}\right). \quad (4.53)$$

证明 令 $G_i = Y_i - RX_i (i=1, 2, \dots, N)$, $g_i = y_i - Rx_i (i=1, 2, \dots, n)$, 则 $G = Y - R\bar{X} = 0$.

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 &= \frac{1}{n-1} \sum_{i=1}^n [(y_i - Rx_i) - (\hat{R} - R)x_i]^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n [g_i - (\hat{R} - R)x_i]^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n g_i^2 + (\hat{R} - R)^2 \frac{1}{n-1} \sum_{i=1}^n x_i^2 \\ &\quad - 2(\hat{R} - R) \frac{\sum_{i=1}^n g_i x_i}{n-1}, \end{aligned} \quad (4.54)$$

而

$$\begin{aligned} E\left[\frac{1}{n-1} \sum_{i=1}^n g_i^2\right] &= E\left[\frac{1}{n-1} \left(\sum_{i=1}^n (g_i - \bar{g})^2 + n\bar{g}^2\right)\right] \\ &= \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 + \frac{n}{n-1} V(\bar{g}) \\ &= \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 + O\left(\frac{1}{n}\right), \end{aligned} \quad (4.55)$$

$$E\left[(\hat{R} - R)^2 \frac{\sum_{i=1}^n x_i^2}{n-1}\right] \leq \frac{n}{n-1} (X^*)^2 E(\hat{R} - R)^2 \\ = O\left(\frac{1}{n}\right). \quad (4.56)$$

其中 $X^* = \max_{1 \leq i \leq N} \{X_i\}$;

$$\left| E\left[(\hat{R} - R) \frac{\sum_{i=1}^n g_i x_i}{n-1}\right] \right| \leq \frac{X^*}{n-1} \left| E\left[(\hat{R} - R) \sum_{i=1}^n g_i\right] \right| \\ = \frac{X^* n}{n-1} |E[(\hat{R} - R) \bar{g}]| \\ \leq \frac{X^* n}{n-1} \sqrt{E(\hat{R} - R)^2} \sqrt{E(\bar{g}^2)} \\ = O\left(\frac{1}{n}\right). \quad (4.57)$$

将(4.54)式取期望, 并将(4.55)~(4.57)各式代入, 即得

$$E\left[\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2\right] = \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 + O\left(\frac{1}{n}\right). \blacksquare$$

推论1 若令

$$v_1(\hat{R}) = \frac{1-f}{n\bar{X}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \quad (4.58)$$

$$= \frac{1-f}{n\bar{X}^2} (S_y^2 + \hat{R}^2 S_x^2 - 2\hat{R}S_{yx}), \quad (4.59)$$

四

$$E[v_1(\hat{R})] = \frac{1-f}{n\bar{X}^2} \cdot \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 + O\left(\frac{1}{n^2}\right) \quad (4.60)$$

$$= \text{MSE}(\hat{R}) + O\left(\frac{1}{n^2}\right) = V(\hat{R}) + O\left(\frac{1}{n^2}\right). \quad (4.61)$$

推论2 若令

$$v_2(\hat{R}) = \frac{1-f}{n\bar{x}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2, \quad (4.62)$$

则也有

$$E[v_2(\hat{R})] = \frac{1}{n\bar{x}^2} \cdot \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 + O\left(\frac{1}{n^2}\right) \quad (4.63)$$

$$= \text{MSE}(\hat{R}) + O\left(\frac{1}{n^2}\right) = V(\hat{R}) + O\left(\frac{1}{n^2}\right). \quad (4.64)$$

证明 $E\left[\frac{1}{\bar{x}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2\right]$

$$\begin{aligned}
&= E \left[\frac{1}{\bar{x}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \right. \\
&\quad \left. - \frac{\bar{x}^2 - \bar{X}^2}{\bar{x}^2 \bar{X}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \right] \\
&= \frac{1}{\bar{X}^2} E \left[\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \right] \\
&\quad - E \left[\frac{\bar{x}^2 - \bar{X}^2}{\bar{x}^2 \bar{X}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \right].
\end{aligned}$$

记 $Y^* = \max_{1 \leq i \leq N} \{Y_i\}$, 于是

$$\begin{aligned}
\sum_{i=1}^n (y_i - \hat{R}x_i)^2 &\leq 2 \sum_{i=1}^n (y_i^2 + \hat{R}^2 x_i^2) \\
&\leq 2n \left[(Y^*)^2 + \left(\frac{Y^*}{X^*} \right)^2 (X^*)^2 \right],
\end{aligned}$$

$$\begin{aligned}
&\left| E \left[\frac{\bar{x}^2 - \bar{X}^2}{\bar{x}^2 \bar{X}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \right] \right| \\
&\leq \frac{2n}{n-1} \left[(Y^*)^2 + \left(\frac{Y^*}{X^*} \right)^2 (X^*)^2 \right] \frac{X^* + \bar{X}}{\bar{X}^2 \bar{X}^2} E \bar{x} - \bar{X} \left[O\left(\frac{1}{n}\right) \right].
\end{aligned}$$

从而

$$\begin{aligned}
E \left[\frac{1}{\bar{x}^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{R}x_i)^2 \right] &= \frac{1}{\bar{X}^2} \cdot \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 \\
&\quad + O\left(\frac{1}{n}\right), \quad (4.65)
\end{aligned}$$

$$E[v_2(\hat{R})] = \frac{1-f}{n\bar{X}^2} \cdot \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 + O\left(\frac{1}{n^2}\right). \blacksquare$$

§ 4.3 分层随机抽样中的比估计

对于分层随机样本, 可以定义两种比估计的方法: 一种是先对各层分别进行比估计, 然后汇总成按层权平均得到总体参数的估计, 这称为分别比估计. 另一种是先按分层随机抽样公式对 \bar{Y} 、 \bar{X} (或 Y 、 X) 作分层估计, 再对它们应用比估计, 称为联合比估计. 本节分别就这两种情况进行讨论, 并对两者进行比较.

4.3.1 分别比估计

定义 4.2 对分层随机抽样, \bar{y}_h, \bar{x}_h 是 h 层样本均值, \bar{y}_{R_h} 与 \hat{Y}_{R_h} 是 h 层 \bar{Y}_h 与 Y_h 的比估计, \bar{X}_h, X_h 分别是 h 层 \mathcal{X} 的均值与总和, 则总体均值 \bar{Y} 与总和 Y 的以下估计称为分别比估计 (separate ratio estimator):

$$\bar{y}_{R_s} = \sum_h W_h \bar{y}_{R_h} = \sum_h W_h \frac{\bar{y}_h}{\bar{x}_h} \bar{X}_h = \frac{1}{N} \sum_h \frac{y_h}{x_h} X_h, \quad (4.66)$$

$$\hat{Y}_{R_s} = N \bar{y}_{R_s} = \sum_h \frac{y_h}{x_h} X_h = \sum_h \hat{Y}_{R_h}. \quad (4.67)$$

为方便起见, 下面仅对 \hat{Y}_{R_s} 进行讨论.

定理 4.6 在分层随机抽样中, 若每层的样本量 n_h 都比较大, 则有

$$E(\hat{Y}_{R_s}) \approx Y, \quad (4.68)$$

$$\begin{aligned} \text{MSE}(\hat{Y}_{R_s}) &\approx V(\hat{Y}_{R_s}) \\ &\approx \sum_h \frac{N_h^2(1-f_h)}{n_h} (S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h \rho_h S_{y_h} S_{x_h}), \end{aligned} \quad (4.69)$$

其中

$$R_h = \frac{Y_h}{X_h} = \frac{\bar{Y}_h}{\bar{X}_h},$$

$$\rho_h = \frac{S_{y_h x_h}}{S_{y_h} S_{x_h}}.$$

证明 根据定理 4.1, 当 n_h 大时, 有

$$E(\hat{Y}_{R_h}) \approx Y_h,$$

$$\begin{aligned} \text{MSE}(\hat{Y}_{R_h}) &\approx V(\hat{Y}_{R_h}) \\ &\approx \frac{N_h^2(1-f_h)}{n_h} (S_{y_h}^2 + R_h^2 S_{x_h}^2 - 2R_h \rho_h S_{y_h} S_{x_h}). \end{aligned}$$

于是根据 \hat{Y}_{R_s} 的定义, 注意到各层的抽样是相互独立的, 从而定理得证. ■

特别指出: 定理 4.6 的条件是每层的 n_h 都要比较大. 当 n_h 不太大, 而层数 L 比较大时 (不少实际问题即是如此), \hat{Y}_{R_s} 的偏倚就可能比较大.

4.3.2 联合比估计

定义 4.3 对分层随机抽样, 总体总和与均值的如下估计称为联合比估计 (combined ratio estimator):

$$\hat{P}_{Ro} = \frac{\bar{y}_{st}}{\bar{x}_{st}} X = \frac{\hat{P}_{st}}{\hat{X}_{st}} X \triangleq \hat{R}_o X, \quad (4.70)$$

$$\bar{y}_{Ro} = \frac{\hat{P}_{Ro}}{N} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} \triangleq \hat{R}_o \bar{X}. \quad (4.71)$$

其中 $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, $\bar{x}_{st} = \sum_h W_h \bar{x}_h$ 与 \hat{P}_{st} , \hat{X}_{st} 分别是 \bar{Y} , \bar{X} 与 Y , X 的分层简单估计, 而 $\hat{R}_o = \hat{P}_{st} / \hat{X}_{st} = \bar{y}_{st} / \bar{x}_{st}$.

注意: 联合比估计只需已知 X (或 \bar{X}), 而不需要已知每层的 X_h .

定理 4.7 对于分层随机抽样, 若总样本量 n 比较大, 则

$$E(\hat{P}_{Ro}) \approx Y, \quad (4.72)$$

$$\begin{aligned} \text{MSE}(\hat{P}_{Ro}) &\approx V(\hat{P}_{Ro}) \\ &\approx \sum_h \frac{N_h^2(1-f_h)}{n_h} (S_{yh}^2 + R^2 S_{xh}^2 - 2R\rho_h S_{yh} S_{xh}). \end{aligned} \quad (4.73)$$

证明 采用与证明定理 4.1 类似的方法. 当 n 大时, $\bar{x}_{st} \approx \bar{X}$.

$$\begin{aligned} \hat{P}_{Ro} - Y &= \frac{\bar{y}_{st}}{\bar{x}_{st}} X - RX \\ &= \frac{N\bar{X}}{\bar{x}_{st}} (y_{st} - R\bar{x}_{st}) \\ &\approx N(\bar{y}_{st} - R\bar{x}_{st}). \end{aligned}$$

于是 $E(\hat{P}_{Ro} - Y) \approx 0$, (4.72) 式成立. 又令

$$G_h = Y_h - R X_h, \quad \text{则 } \bar{G}_h = \bar{Y}_h - R \bar{X}_h.$$

$$\bar{g}_{st} = \bar{y}_{st} - R\bar{x}_{st}, \quad E(\bar{g}_{st}) = \bar{Y} - R\bar{X} = \bar{G} = 0.$$

于是

$$\begin{aligned} V(\hat{P}_{Ro}) &\approx E(\hat{P}_{Ro} - Y)^2 \approx N^2 E(\bar{g}_{st}^2) = N^2 V(\bar{g}_{st}) \\ &= \sum_h \frac{N_h(N_h - n_h)}{n_h} S_{gh}^2. \end{aligned}$$

其中

$$\begin{aligned} S_{gh}^2 &= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (G_{hi} - \bar{G}_h)^2 \\ &= \frac{1}{N_h - 1} \sum_{i=1}^{N_h} [(Y_{hi} - \bar{Y}_h) - R(X_{hi} - \bar{X}_h)]^2 \\ &= S_{yh}^2 + R^2 S_{xh}^2 - 2R\rho_h S_{yh} S_{xh}. \end{aligned} \quad (4.74)$$

从而定理得证. ■

比较(4.69)与(4.73)两式可知 $V(\hat{P}_{Rt})$ 与 $V(\hat{P}_{Ro})$ 两个近似公式形式上非常相像, 所不同的是, 前者的和式中出现的是各层的 R_h ; 而后者的和式中出现的是总体的 R .

用 Hartley-Ross 方法可得到 \hat{P}_{RC} 的偏倚与它的标准差之比的上界。由

$$\begin{aligned}\text{Cov}(\hat{R}_c, \bar{x}_{st}) &\triangleq \text{Cov}\left(\frac{\bar{y}_{st}}{\bar{x}_{st}}, \bar{x}_{st}\right) \\ &= E(\bar{y}_{st}) - E(\hat{R}_c)E(\bar{x}_{st}) \\ &= Y - \bar{X}E(\hat{R}_c).\end{aligned}\quad (4.75)$$

$$\begin{aligned}|E(\hat{R}_c) - R| &= \frac{1}{\bar{X}} |\text{Cov}(\hat{R}_c, \bar{x}_{st})| \\ &\leq \frac{1}{\bar{X}} \sqrt{V(\hat{R}_c)} \cdot \sqrt{V(\bar{x}_{st})} \\ &= \sqrt{V(\hat{R}_c)} \text{Cv}(\bar{x}_{st}),\end{aligned}$$

从而

$$\frac{|E(\hat{P}_{RC}) - Y|}{\sqrt{V(\hat{P}_{RC})}} = \frac{|E(\hat{R}_c) - R|}{\sqrt{V(\hat{R}_c)}} \leq \text{Cv}(\bar{x}_{st}). \quad (4.76)$$

4.3.3 分别比估计与联合比估计的比较

当每层的 n_h 都比较大时(此时 n 更大), 根据定理 4.6 与定理 4.7, 有

$$\begin{aligned}V(\hat{P}_{RC}) - V(\hat{P}_{RS}) &\approx \sum_h \frac{N_h^2(1-f_h)}{n_h} [(R^2 - R_h^2)S_{xh}^2 - 2(R - R_h)\rho_h S_{yxh}S_{xh}] \\ &= \sum_h \frac{N_h^2(1-f_h)}{n_h} [(R - R_h)^2 S_{xh}^2 - 2(R - R_h) \\ &\quad \times (\rho_h S_{yxh}S_{xh} - R_h S_{xh}^2)].\end{aligned}\quad (4.77)$$

当每层的 Y_h 与 X_h 的关系是通过原点的直线关系, 也即 Y_h 与 X_h 成正比例时, $\rho_h S_{yxh} = R_h S_{xh}^2$, 此时(4.77)式中括号中的第二项为 0. 一般的只要比估计有效($\rho_h > \frac{1}{2} \frac{C_{hx}}{C_{hy}}$), 则这一项的值不会很大, 因而除非各层的 $R_h = R$, 否则, 有

$$V(\hat{P}_{RC}) > V(\hat{P}_{RS}). \quad (4.78)$$

因此只要各层的 n_h 都比较大, 各层的比估计比较有效, 则分别估计要优于联合估计. 但当某些 n_h 不太大时, 则应用联合估计, 因为此时分别估计的偏倚可能很大, 从而使总的均方误差增大.

对上述的近似方差作估计时, 可用 \hat{R}_h, \hat{R}_c 估计 R_h, R ; 用 s_{yh}^2, s_{xh}^2 估计 S_{yh}^2, S_{xh}^2 ; 用样本协方差 s_{yxh} 估计层协方差 $S_{yxh} = \rho_h S_{yh}S_{xh}$.

4.3.4 分层比估计时的最优分配

在分层抽样中,若用比估计,则最优分配与用简单估计时的最优分配稍有不同。下面以分别比估计 \hat{Y}_{Rr} 为例说明考虑的方法(联合比估计也有类似的结果)。

根据定理 4.6, 当 n_h 大时, 有

$$\begin{aligned} V(\hat{Y}_{Rr}) &\approx \sum_h \frac{N_h(N_h - n_h)}{n_h} (S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h \rho_{yx} S_{yh} S_{xh}) \\ &\triangleq \sum_h \frac{N_h(N_h - n_h)}{n_h} S_{yh}^2. \end{aligned} \quad (4.79)$$

其中

$$S_{yh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} y_{hi}^2 \triangleq \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - R_h X_{hi})^2. \quad (4.80)$$

用 § 3.3 同样的处理方法, 可得到在给定总费用为简单线性函数

$$C = c_0 + \sum_h c_h n_h$$

时, 最优分配为

$$n_h \propto \frac{N_h S_{yh}}{\sqrt{c_h}}. \quad (4.81)$$

这里的主要问题是: 各层 S_{yh} 的值难于确定, 在多数场合需要凭经验。在一些问题中, S_{yh} 可以看成与 \bar{X}_h 或 $\sqrt{\bar{X}_h}$ 近似成正比。

4.3.5 数值例子——耕地面积核实调查

例 4.3 为核实某地区上报耕地面积数字的真实性, 对该地区所属的村按不同的地形面貌划分为三个层, 用比例分配分层随机抽样在层内抽取村进行耕地面积核查。结果如表 4.4 所示, 其中 y_i 是第 i 个样本村实际耕地面积, x_i 是该村登记在册的耕地面积。

一些有关层的参数及根据表 4.4 计算的一些中间结果列于表 4.5。

表 4.5 中的 \bar{y} 、 \bar{x} 是 $n = 23$ 个村样本均值, $\hat{R} = \bar{y}/\bar{x}$ 。由于本例中各层的分配是比例分配, 因此, 这也可看作是从总体中抽取的一个简单随机样本的均值和对总体 $R = Y/X$ 的估计。而 s_y^2 、 s_x^2 、 s_{yx} 分别是这个样本关于 y_i 、 x_i 的样本方差和样本协方差。

我们按不同抽样方法及采用不同的估计量来对该地区实际耕地面积 Y 作出估计, 并给出估计的精度。所有的 fpc 都取为 1。

表 4.4 耕地面积核实调查样本数据

h=1			h=2			h=3		
i	y _i	x _i	i	y _i	x _i	i	y _i	x _i
1	1241	1174	1	1030	885	1	652	527
2	858	945	2	931	996	2	627	585
3	961	884	3	1039	805	3	974	741
4	1132	1113	4	1101	995	4	1499	1130
5	934	1031	5	941	831	5	1200	1140
6	838	792	6	561	545	6	1254	952
7	621	586	7	930	807			
8	647	609						
9	654	599						
10	848	827						

表 4.5 耕地核查各层参数及若干中间结果

h	N_h	W_h	X_h	n_h	y_h	\bar{x}_h	$\hat{E}_h = y_h/x_h$
1	427	0.4375	367200	10	873 4000	856.0000	1.020327
2	297	0.3043	251600	7	933 2857	837.7143	1.114086
3	252	0.2582	208000	6	1034 3333	845.8333	1.222857
N=976			N=268000	n=23	$\bar{y}=933.6087$	$\bar{x}=847.7826$	$\hat{E}=1.101236$

h	$s_{y h}^2$	$s_{x h}^2$	$s_{yx h}$	$Q_h = W_h^2/n_h$	V_h	V_h^*
1	42064.933	45710.889	42055.667	0.0191406	3831.9745	4864.0154
2	31221.571	23294.905	22692.429	0.0132283	9572.2837	9489.1195
3	121470.667	71846.967	87713.867	0.0111112	1485.951	15423.045
$s_y^2=57745$		$s_x^2=41446$	$s_{yx}=43051$			

1) 简单随机抽样, 简单估计

$$\hat{P} = N\bar{y} = 976 \times 933.6087 = 911202,$$

$$s(\hat{P}) = \sqrt{v(\hat{P})} = N\sqrt{v(\bar{y})}.$$

为计算 $v(\bar{y})$, 我们采用 (3.53) 式:

$$v(\bar{y}) \approx \frac{1}{n} \left[\frac{n-1}{n} s_y^2 + v(\bar{y}_{st}) \right].$$

其中

$$s_y^2 = 57745,$$

$$v(\bar{y}_{st}) = \frac{1}{n} \sum_h W_h s_{y|h}^2 = 2576.86,$$

从而
$$s(\hat{P}) = 976 \times \sqrt{\frac{1}{23} \left[\frac{22}{23} \times 57745 + 2576.86 \right]}$$

$$= 48932.$$

2) 简单随机抽样, 比估计

$$\hat{P}_R = \frac{\bar{y}}{\bar{x}} X = \hat{R} X = 1.101236 \times 826800 = 910502,$$

$$s(\hat{P}_R) = N \cdot \sqrt{\frac{1}{n} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{yx})}$$

$$= 976 \times \left[\frac{1}{23} (57745 + 1.101236^2 \times 41446 \right. \\ \left. - 2 \times 1.101236 \times 43051) \right]^{\frac{1}{2}}$$

$$= 23372.$$

这里在利用分层样本估计总体方差与协方差时, 用了近似公式. 直接用样本方差 s_y^2 、 s_x^2 与协方差 s_{yx} 进行估计(从 $v(\bar{y})$ 的计算过程可看出两者相差不大).

3) 分层随机抽样, 简单估计

$$\hat{P}_{st} = \sum_k N_k \bar{y}_k = 910780,$$

$$s(\hat{P}_{st}) = N \sqrt{v(\bar{y}_{st})} = 976 \times \sqrt{2576.86} = 49544.$$

4) 分层随机抽样, 分别比估计

$$\hat{P}_{Rs} = \sum_k \frac{\bar{y}_k}{\bar{x}_k} X_k = \sum_k \hat{R}_k X_k = 909459,$$

$$s(\hat{P}_{Rs}) = N \sqrt{\sum_k \frac{W_k^2}{n_k} (s_{y_k}^2 + \hat{R}_k^2 s_{x_k}^2 - 2\hat{R}_k s_{y_k x_k})}$$

$$\triangleq N \sqrt{\sum_k Q_k V_k}$$

$$= 976 \times (0.0191406 \times 3831.9745 + 0.0132283 \\ \times 9572.2837 + 0.0111112 \times 1485.951)^{\frac{1}{2}}$$

$$= 14360.$$

5) 联合比估计

$$\hat{P}_{st} = 910780, \quad \hat{X}_{st} = \sum N_k \bar{x}_k = 827463,$$

$$\hat{R}_c = \frac{\hat{P}_{st}}{\hat{X}_{st}} = 1.100690, \quad \hat{P}_{Ro} = \hat{R}_c X = 910050,$$

$$s(\hat{P}_{Ro}) = N \sqrt{\sum_k \frac{W_k^2}{n_k} (s_{y_k}^2 + \hat{R}_c^2 s_{x_k}^2 - 2\hat{R}_c s_{y_k x_k})}$$

$$\begin{aligned}
& \triangleq N \sqrt{\sum_k Q_k V_k} \\
& = 976(0.0191406 \times 4864.0154 + 0.0132283 \\
& \quad \times 9489.1195 + 0.0111112 \times 15423.045)^{\frac{1}{2}} \\
& = 19274.
\end{aligned}$$

上述五种结果可以列表进行比较(见表4.6)。其中相对精度是每种方法所得估计量的方差与简单随机抽样的简单估计的方差之比的倒数(相当于deff的倒数)。从表4.6中可以看出在五种估计中,以分层随机抽样的两种比估计的精度最高,其中分别比估计的方差比联合比估计更小。但我们已指出,当 n_h 不大时,考虑到偏倚,联合比估计不一定比分别比估计效果差。此外,简单随机抽样的比估计效果也不错。在此例中,分层随机抽样的简单估计效果不好。这是因为层内方差较大(特别是 $h=3$),甚至超过了总体方差的缘故。表4.6中 \bar{Y} 的五种估计差别不很大,这纯属偶然,是因为正好抽到一个“较好”的样本。按标准差值看,两个简单估计的变化幅度较大。

表4.6 不同抽样方法和估计方法的比较

抽样方法	估计方法	估计量 \hat{Y}	\hat{Y} 的标准差估计 $s(\hat{Y})$	相对精度
简单随机抽样	简单估计	911202	48932	1.00
简单随机抽样	比估计	910102	23372	4.38
分层随机抽样	简单估计	910780	49544	0.98
分层随机抽样	分别比估计	909459	14360	11.61
分层随机抽样	联合比估计	910050	19274	6.45

§ 4.4 消除或减少比估计偏倚的方法

由于比估计是有偏的,因此当样本量不太大时,特别是在分层随机抽样中,若层数很大,而每层样本量 n_h 不大且分别比估计又适用的情形,比估计量的偏倚就可能很大。此时可引进无偏的比类型估计量或设法减少估计量的偏倚(例如使它的阶从 $O\left(\frac{1}{n}\right)$ 降至 $O\left(\frac{1}{n^2}\right)$ 或更小)的办法来处理。因此这类估计量是有实用意义的。不过也应指出,在消除偏倚或减少偏倚的同时,可能引起估计量方差的增加。因此要全面衡量这些新的估计量,还应研究这些估计量的总的均方误差,另外还应考虑这些估计量能否提供一种从样本估计其方差(或近似方差)的可行方法。

4.4.1 无偏的比类型估计量

本段讨论三种无偏的比类型的估计量:

一、Hartley-Ross(1954)估计量

考虑总体 $R = \frac{\bar{Y}}{\bar{X}}$ 的以下估计:

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n r_i \triangleq \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}, \quad (4.82)$$

为求 \bar{r} 的偏倚, 注意到

$$\begin{aligned} E(\bar{r}) &= \frac{1}{N} \sum_{i=1}^N R_i \triangleq \frac{1}{N} \sum_{i=1}^N \frac{Y_i}{X_i} = \bar{R}, \\ \frac{1}{N} \sum_{i=1}^N R_i (X_i - \bar{X}) &= \frac{1}{N} \sum_{i=1}^N Y_i - \frac{\bar{X}}{N} \sum_{i=1}^N R_i = \bar{Y} - \bar{X} \bar{R}. \end{aligned}$$

因而

$$\begin{aligned} E(\bar{r}) - R &= \bar{R} - R \\ &= -\frac{1}{\bar{X}N} \sum_{i=1}^N R_i (X_i - \bar{X}) \\ &= -\frac{1}{\bar{X}N} \sum_{i=1}^N (R_i - \bar{R})(X_i - \bar{X}). \end{aligned}$$

另一方面, 注意到

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^n (r_i - \bar{r})(x_i - \bar{x}) &= \frac{1}{n-1} \sum_{i=1}^n r_i (x_i - \bar{x}) \\ &= \frac{n}{n-1} (\bar{y} - \bar{r}\bar{x}) \end{aligned}$$

$$\text{是} \quad \frac{1}{N-1} \sum_{i=1}^N (R_i - \bar{R})(X_i - \bar{X}) = \frac{1}{N-1} \sum_{i=1}^N R_i (X_i - \bar{X})$$

的一个无偏估计. 因此

$$\hat{R}_{HR} = \bar{r} + \frac{n(N-1)}{(n-1)\bar{X}N} (\bar{y} - \bar{r}\bar{x}) \quad (4.83)$$

是 R 的一个无偏估计.

二、Mickey(1959)估计量

令

$$\hat{R}_{-i} = \frac{1}{n} \sum_{j=1}^n \hat{R}_{-i,j} = \frac{1}{n} \sum_{j=1}^n \frac{\bar{y}_{-i,j}}{\bar{x}_{-i,j}}, \quad (4.84)$$

其中 \hat{R}_{-i} 是在 n 个样本数据中去掉第 i 个, 其余 $n-1$ 个 y_j 的平均值 \bar{y}_{-i} 与 x_j 的平均值 \bar{x}_{-i} 之比. 用 \hat{R}_{-i} 代替 (4.82) 中的 \bar{r} , 则 Mickey 估计量

$$\hat{R}_x = \hat{R}_- + \frac{n(N-n+1)}{N\bar{X}}(y - \hat{R}_-\bar{x}) \quad (4.85)$$

也是 R 的无偏估计.

三、Lahiri(1951)估计量

Lahiri 估计量是基于一种不等概率抽样所得的样本, 如果每一特定的样本被抽到的概率与样本中辅助变量指标和 $\sum_{i=1}^n x_i$ 成正比, 则按通常意义的比估计量 $\hat{R}_L = \frac{\bar{y}}{\bar{x}}$ 是 R 的无偏估计.

我们先给出两种简单的可以满足上述要求的抽样方法, 更详尽的讨论见第五章.

1. Lahiri (1951) 方法 令 T 为总体中最大的 n 个 X_i 的和, 抽一个 $[0, T]$ 范围内的随机数 ν 及按简单随机抽样抽取 n 个单元, 若这 n 个单元中的 x_i 之和 $\sum_{i=1}^n x_i \geq \nu$, 则这 n 个单元就作为抽中的样本; 否则, 抛弃这 n 个单元, 重抽随机数 ν 及 n 个单元, 再按前面的原则判断所抽的单元留作样本还是再舍弃重抽. 显然, 这样得到的样本是符合要求的.

2. 水野 (Midzuno, 1951) 方法 按与 X_i 成正比的概率在总体中抽取一个单元入样, 再按简单随机抽样在剩下的 $N-1$ 个单元中抽取一个样本量为 $n-1$ 的样本, 则两者组成的样本 S 即是满足要求的样本.

下面我们证明 S 被抽中的概率 P_S 与 $\sum_{i=1}^N x_i$ 成正比.

设 i 是 S 中第一个被抽中的单元, 且抽中 S 的概率为

$$\frac{x_i}{X} \binom{N-1}{n-1}^{-1},$$

对于样本 S 中的每一个单元, 都有可能在第一次抽样中就被抽到, 因此抽中 S 的总概率为

$$P_S \triangleq P_r(S) = \sum_{i=1}^N \frac{x_i}{X \binom{N-1}{n-1}} \propto \sum_{i=1}^N x_i. \quad (4.86)$$

按这种抽样方法, Lahiri 估计量

$$\hat{R}_L = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{\bar{y}}{\bar{x}} \quad (4.87)$$

的期望值为:

$$\begin{aligned}
 E(\hat{R}_L) &= \sum_s \left(P_s \frac{\sum_i y_i}{\sum_i x_i} \right) = \sum_s \frac{\sum_i x_i}{X \binom{N-1}{n-1}} \cdot \frac{\sum_i y_i}{\sum_i x_i} \\
 &= \frac{1}{X \binom{N-1}{n-1}} \sum_s \left(\sum_{i=1}^n y_i \right) \\
 &= \frac{1}{X \binom{N-1}{n-1}} \binom{N}{n} \frac{n}{N} \sum_{i=1}^N Y_i = R.
 \end{aligned}$$

从而 \hat{R}_L 是无偏的.

4.4.2 减少比估计偏倚的方法

在 4.2.2 中指出 \hat{R} 的偏倚的阶为 $O\left(\frac{1}{n}\right)$, 有一些改进估计的方法可使它的阶降低到 $O\left(\frac{1}{n^2}\right)$ 或更小.

一、Jackknife 方法

Durbin (1954) 首先将 Quenouille 的 Jackknife 方法用于比估计, 从而降低其偏倚的阶.

先不考虑 fpc 的影响, 即若 $1-f = \frac{N-n}{N} \approx 1$ 时, 记

$$E(\hat{R}) = R + \frac{b_1}{n} + \frac{b_2}{n^2} + O\left(\frac{1}{n^3}\right). \quad (4.88)$$

设 $n = mg$, 将样本随机地分为 g 组, 每组 (子样本) 的大小为 m , 则

$$E(g\hat{R}) = gR + \frac{b_1}{m} + \frac{b_2}{gm^2} + O\left(\frac{1}{n^2m}\right). \quad (4.89)$$

令 \hat{R}_{-j} 是在样本中舍弃第 j 组数据后求得的比估计, 由于此时样本量为 $m(g-1)$, 因此

$$\begin{aligned}
 E(\hat{R}_{-j}) &= R + \frac{b_1}{(g-1)m} + \frac{b_2}{(g-1)^2m^2} + O\left(\frac{1}{n^2m}\right), \\
 E[(g-1)\hat{R}_{-j}] &= (g-1)R + \frac{b_1}{m} + \frac{b_2}{(g-1)m^2} + O\left(\frac{1}{n^2m}\right).
 \end{aligned} \quad (4.90)$$

将 (4.89) 与 (4.90) 两式相减, 则得

$$E[g\hat{R} - (g-1)\hat{R}_{-j}] = R - \frac{b_2}{g(g-1)m^2} + O\left(\frac{1}{n^2m}\right)$$

$$= R + O\left(\frac{1}{n^2}\right).$$

R 的 Jackknife 估计即是

$$\hat{R}_J \triangleq g\hat{R} - (g-1) \left[\frac{1}{g} \sum_{j=1}^g \hat{R}_{-j} \right] \quad (4.91)$$

$$\triangleq g\hat{R} - (g-1)\hat{R}_-, \quad (4.92)$$

其中 \hat{R}_- 是 g 组 \hat{R}_{-j} 的平均数. 显然

$$E(\hat{R}_J) = R + O\left(\frac{1}{n^2}\right). \quad (4.93)$$

最常用且简单的情形是 $m=1$, $g=n$, 即每次舍弃一个样本数据, 此时有

$$\hat{R}_J = n\hat{R} - (n-1)\hat{R}_-, \quad (4.94)$$

若 $1-f$ 不能忽略, 则 \hat{R} 的偏倚可表成

$$\begin{aligned} E(\hat{R} - R) &= \frac{b_1(1-f)}{n} + O\left(\frac{1}{n^2}\right) \\ &= \frac{b_1}{n} - \frac{b_1}{N} + O\left(\frac{1}{n^2}\right). \end{aligned}$$

为了消除 $\frac{b_1}{n} - \frac{b_1}{N}$, 按上述随机分组的步骤, 令

$$\hat{R}'_J = \omega\hat{R} - (\omega-1)\hat{R}_-, \quad (4.95)$$

其中

$$\omega = g\left(1 - \frac{n-m}{N}\right). \quad (4.96)$$

则

$$E(\hat{R}'_J) = R + O\left(\frac{1}{n^2}\right).$$

证明留作练习.

二、Beale(1962)估计量

$$\hat{R}_B = \frac{\hat{y} + \frac{1-f}{n} \cdot \frac{s_{yx}}{\bar{x}}}{\bar{x} + \frac{1-f}{n} \cdot \frac{s_x^2}{\bar{x}}} = \hat{R} \cdot \frac{1 + \frac{1-f}{n} c_{yx}}{1 + \frac{1-f}{n} c_x^2}. \quad (4.97)$$

三、Tin(1965)估计量

$$\hat{R}_T = \hat{R} \left[1 - \frac{1-f}{n} \left(\frac{s_x^2}{\bar{x}^2} - \frac{s_{yx}}{\bar{y}\bar{x}} \right) \right] = \hat{R} \left[1 - \frac{1-f}{n} (c_x^2 - c_{yx}) \right]. \quad (4.98)$$

这两个估计量偏倚的阶皆为 $O\left(\frac{1}{n^2}\right)$.

§ 4.5 回归估计量(β 设定时的情形)

4.5.1 回归估计量的一般形式

若 Y_i 对 X_i 的回归直线不通过原点, 则为了提高估计精度, 可进一步用回归估计量(regression estimator)来代替比估计量.

定义 4.4 对于简单随机抽样, 总体均值 \bar{Y} 与总和 Y 的(线性)回归估计量定义为

$$\bar{y}_{lr} = \bar{y} + \beta(\bar{X} - \bar{x}) = \bar{y} - \beta(\bar{x} - \bar{X}), \quad (4.99)$$

$$\hat{Y}_{lr} = N\bar{y}_{lr}. \quad (4.100)$$

其中 \bar{y} 、 \bar{x} 是样本均值, β 可以是事先设定的常数, 也可以是从样本中计算得到的某一特定的统计量, 例如样本回归系数.

简单估计量与比估计量都可以看作是上面一般情形的回归估计量的特殊情形. 在(4.99)式中, 若令 $\beta = 0$, 则 $\bar{y}_{lr} = \bar{y}$ 即是简单估计量; 若令 $\beta = \frac{\bar{y}}{\bar{x}}$, 则 $\bar{y}_{lr} = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \hat{y}_R$ 即是比估计量.

4.5.2 β 设定情形的一般结果

关于 β 为事先设定时的回归估计量的性质, 讨论起来, 比较简单. 在许多实际问题中, 也确实可以将 β 事先给定, 例如为同样目的进行的调查若已重复多次, 则有理由将从已往的资料中得出的 Y_i 对 X_i 的回归系数作为 β 的设定值.

定理 4.8 若 β_0 是设定常数, 则回归估计量

$$\bar{y}_{lr} = \bar{y} + \beta_0(\bar{X} - \bar{x}) \quad (4.101)$$

是 \bar{Y} 的无偏估计, 且

$$\begin{aligned} V(\bar{y}_{lr}) &= \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N [(Y_i - \bar{Y}) + \beta_0(X_i - \bar{X})]^2 \\ &= \frac{1-f}{n} (S_y^2 + \beta_0^2 S_x^2 - 2\beta_0 S_{yx}). \end{aligned} \quad (4.102)$$

注意: 在定理中并没有对 \mathscr{Y} 与 \mathscr{X} 的关系作任何假定.

证明 $E(\bar{y}_{lr}) = E(\bar{y}) + \beta_0 E(\bar{X} - \bar{x}) = \bar{Y}$,

而 \bar{y}_{lr} 又可表为 $y_i + \beta_0(\bar{X} - x_i)$ 的样本均值. 后者的总体均值为 \bar{Y} , 故根据定理 2.2, 有

$$\begin{aligned}
 V(\bar{y}_{lr}) &= \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N [(Y_i - \bar{Y}) + \beta_0(\bar{X} - X_i)]^2 \\
 &= \frac{1-f}{n} \frac{1}{N-1} \sum_{i=1}^N [(Y_i - \bar{Y}) - \beta_0(X_i - \bar{X})]^2 \\
 &= \frac{1-f}{n} (S_y^2 + \beta_0^2 S_x^2 - 2\beta_0 S_{yx}). \blacksquare
 \end{aligned}$$

推论 若 s_y^2, s_x^2, s_{yx} 分别是简单随机样本的方差与协方差, 则

$$v(y_{lr}) = \frac{1-f}{n} (s_y^2 + \beta_0^2 s_x^2 - 2\beta_0 s_{yx}) \quad (4.103)$$

是 $V(\bar{y}_{lr})$ 的无偏估计.

定理 4.9 极小化 $V(y_{lr})$ 的 β_0 值为

$$B \triangleq \frac{S_{yx}}{S_x^2} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2} = \frac{\rho S_y}{S_x}, \quad (4.104)$$

且

$$V_{\min}(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1 - \rho^2). \quad (4.105)$$

其中 B 是有限总体情形 \mathscr{Y} 对 \mathscr{X} 的(线性)回归系数, ρ 是相关系数.

证明 在(4.102)式中, 令

$$\beta_0 \triangleq B + \Delta B = \frac{S_{yx}}{S_x^2} + \Delta B,$$

则

$$\begin{aligned}
 V(\bar{y}_{lr}) &= \frac{1-f}{n} \left[S_y^2 + S_x^2 \left(\frac{S_{yx}^2}{S_x^4} + (\Delta B)^2 + 2(\Delta B) \frac{S_{yx}}{S_x^2} \right) \right. \\
 &\quad \left. - 2S_{yx} \left(\frac{S_{yx}}{S_x^2} + \Delta B \right) \right] \\
 &= \frac{1-f}{n} \left[\left(S_y^2 - \frac{S_{yx}^2}{S_x^2} \right) + (\Delta B)^2 S_x^2 \right]. \quad (4.106)
 \end{aligned}$$

它在 $\Delta B = 0$ 即 $\beta_0 = B$ 时, 达到极小值:

$$\begin{aligned}
 V_{\min}(\bar{y}_{lr}) &= \frac{1-f}{n} \left(S_y^2 - \frac{S_{yx}^2}{S_x^2} \right) \\
 &= \frac{1-f}{n} S_y^2 (1 - \rho^2). \blacksquare
 \end{aligned}$$

当 $\beta_0 \neq B$ 时, 根据(4.106)式, 有

$$\begin{aligned}
 V(\bar{y}_{lr}) &= V_{\min}(\bar{y}_{lr}) + \frac{1-f}{n} (\beta_0 - B)^2 S_x^2 \\
 &= V_{\min}(\bar{y}_{lr}) \left[1 + \frac{(\beta_0 - B)^2 S_x^2}{(1 - \rho^2) S_y^2} \right]
 \end{aligned}$$

$$= V_{\min}(\bar{y}_{lr}) \left[1 + \left(\frac{\beta_0}{B} - 1 \right)^2 \frac{\rho^2}{1 - \rho^2} \right].$$

因此,为使方差的相对增加不超过 K ,即

$$\frac{V(\bar{y}_{lr}) - V_{\min}(\bar{y}_{lr})}{V_{\min}(\bar{y}_{lr})} \leq K, \quad (4.107)$$

则必须

$$\left| \frac{\beta_0}{B} - 1 \right| \leq \sqrt{\frac{K(1 - \rho^2)}{\rho^2}}. \quad (4.108)$$

例如若 $\rho = 0.7$, $K = 10\%$, 则

$$\left| \frac{\beta_0}{B} - 1 \right| \leq \sqrt{\frac{0.1 \times 0.51}{0.49}} = 0.32.$$

(4.108)式表明,为保证 $V(\bar{y}_{lr})$ 不会有显著的增加,当 $|\rho|$ 很大时, $\frac{\beta_0}{B}$ 应很接近于 1,也即 β_0 应尽可能接近 B ;而当 ρ 不是很大时,可容许 β_0 偏离 B 稍大些.

4.5.3 差估计量

定义 4.5 $\beta = 1$ 时的回归估计量

$$\bar{y}_d = \bar{y} + (\bar{X} - \bar{x}) = \bar{X} + (\bar{y} - \bar{x}) \triangleq \bar{X} + \bar{d} \quad (4.109)$$

也称为差估计量(difference estimator). 其中

$$\bar{d} = \bar{y} - \bar{x} \quad (4.110)$$

是 $d_i = y_i - x_i$ 的样本均值.

作为 β_0 设定的一种回归估计量, \bar{y}_d 是 \bar{Y} 的无偏估计,且

$$V(\bar{y}_d) = \frac{1-f}{n} (S_y^2 + S_x^2 - 2S_{yx}). \quad (4.111)$$

当 X_i 是调查指标最近一次普查数据时,常可采用差估计量来估计 \bar{Y} (或 Y).

§ 4.6 回归估计量(β 取样本回归系数的情形)

4.6.1 表达式及若干引理

在回归估计量的一般形式(4.99)中,若 β 需根据样本确定,则定理 4.9 表明:一个有效的估计是总体回归系数 B 的最小二乘估计,也即样本回归系数:

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (4.112)$$

此时总体均值 \bar{Y} 的回归估计量为

$$\bar{y}_{lr} = \bar{y} + b(\bar{X} - \bar{x}) = \bar{y} - b(\bar{x} - \bar{X}). \quad (4.113)$$

与比估计量的情况类似, 此时 \bar{y}_{lr} 是有偏的. 为了深入研究它的性质, 在本段中我们先给出若干预备引理.

引理 4.2 若 B 是有限总体中 Y_i 对 X_i 的回归系数, \bar{x} 是从总体中抽取的简单随机样本 x_i 的均值, 又

$$s_i = (Y_i - \bar{Y}) - B(X_i - \bar{X}), \quad (4.114)$$

则

$$1) \sum_{i=1}^N s_i = 0, \quad \bar{s} = 0; \quad (4.115)$$

$$2) \sum_{i=1}^N s_i (X_i - \bar{X}) = 0; \quad (4.116)$$

$$3) E \left[\frac{1}{n} \sum_{i=1}^n s_i (x_i - \bar{x}) \right]^2 = O\left(\frac{1}{n}\right); \quad (4.117)$$

$$4) E \left[\frac{1}{n} \sum_{i=1}^n s_i (x_i - \bar{x}) \right]^4 = O\left(\frac{1}{n^2}\right). \quad (4.118)$$

证明 1) 与 2) 是显然的.

3) 令 $U_i = s_i (X_i - \bar{X})$, 则 $\bar{U} = 0$.

$$\begin{aligned} \left[\frac{1}{n} \sum_{i=1}^n s_i (x_i - \bar{x}) \right]^2 &= \left[\frac{1}{n} \sum_{i=1}^n s_i (x_i - \bar{X}) - \frac{1}{n} \sum_{i=1}^n s_i (\bar{x} - \bar{X}) \right]^2 \\ &= [\bar{u} - \bar{s}(\bar{x} - \bar{X})]^2 \\ &= \bar{u}^2 + \bar{s}^2(\bar{x} - \bar{X})^2 - 2\bar{u}\bar{s}(\bar{x} - \bar{X}). \end{aligned}$$

于是根据引理 4.1, 有

$$E(\bar{u}^2) = O\left(\frac{1}{n}\right),$$

$$E[\bar{s}^2(\bar{x} - \bar{X})^2] = O\left(\frac{1}{n^2}\right),$$

$$E[\bar{u}\bar{s}(\bar{x} - \bar{X})] \leq \sqrt{E(\bar{u}^2)} \sqrt{E[\bar{s}^2(\bar{x} - \bar{X})^2]} = O\left(\frac{1}{n^{3/2}}\right).$$

从而

$$E \left[\frac{1}{n} \sum_{i=1}^n s_i (x_i - \bar{x}) \right]^2 = O\left(\frac{1}{n}\right).$$

$$4) \left[\frac{1}{n} \sum_{i=1}^n s_i (x_i - \bar{x}) \right]^4 = [\bar{u} - \bar{s}(\bar{x} - \bar{X})]^4$$

$$= \bar{u}^4 - 3\bar{u}^3\bar{\varepsilon}(\bar{x} - \bar{X}) + 6\bar{u}^2\bar{\varepsilon}^2(\bar{x} - \bar{X})^2 - 4\bar{u}\bar{\varepsilon}^3(\bar{x} - \bar{X})^3 + \bar{\varepsilon}^4(\bar{x} - \bar{X})^4.$$

而
$$E(\bar{u}^4) = O\left(\frac{1}{n^2}\right),$$

$$E[\bar{u}^3\bar{\varepsilon}(\bar{x} - \bar{X})] \leq \sqrt{E(\bar{u}^6)E[\varepsilon^2(\bar{x} - \bar{X})^2]} = O\left(\frac{1}{n^{5/2}}\right),$$

$$E[\bar{u}^2\bar{\varepsilon}^2(\bar{x} - \bar{X})^2] \leq \sqrt{E(\bar{u}^4)E[\varepsilon^4(\bar{x} - \bar{X})^4]} = O\left(\frac{1}{n^3}\right),$$

$$E[\bar{u}\bar{\varepsilon}^3(\bar{x} - \bar{X})^3] \leq \sqrt{E(\bar{u}^2)E[\varepsilon^6(\bar{x} - \bar{X})^6]} = O\left(\frac{1}{n^{7/2}}\right),$$

$$E[\varepsilon^4(\bar{x} - \bar{X})^4] = O\left(\frac{1}{n^4}\right).$$

从而
$$E\left[\frac{1}{n} \sum_{i=1}^n s_i(x_i - \bar{x})\right]^4 = O\left(\frac{1}{n^2}\right). \quad \blacksquare$$

引理 4.3 B 是有限总体 Y_i 对 X_i 的回归系数, b 是抽自该总体的简单随机样本 y_i 对 x_i 的样本回归系数, $s_i = (y_i - \bar{Y}) - B(x_i - \bar{X})$, 则

$$\begin{aligned} 1) \quad b &= B + \frac{\sum_{i=1}^n s_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= B + \frac{\sum_{i=1}^n s_i(x_i - \bar{X}) - n\bar{s}(\bar{x} - \bar{X})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \end{aligned} \quad (4.119)$$

$$2) \quad E(b - B)^2 = O\left(\frac{1}{n}\right); \quad (4.120)$$

$$3) \quad E(b - B)^4 = O\left(\frac{1}{n^2}\right); \quad (4.121)$$

$$4) \quad E(b) - B = O\left(\frac{1}{\sqrt{n}}\right). \quad (4.122)$$

证明 1)
$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) &= \sum_i y_i(x_i - \bar{x}) \\ &= \sum_i [\bar{Y} + B(x_i - \bar{X}) + s_i](x_i - \bar{x}) \\ &= B \sum_i x_i(x_i - \bar{x}) + \sum_i s_i(x_i - \bar{x}) \\ &= B \sum_i (x_i - \bar{x})^2 + \sum_i s_i(x_i - \bar{X}) - n\bar{s}(\bar{x} - \bar{X}). \end{aligned}$$

从而

$$\begin{aligned} b &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= B + \frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= B + \frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{X})}{\sum_{i=1}^n (x_i - \bar{x})^2} - n\bar{\varepsilon} \frac{\bar{x} - \bar{X}}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned}$$

2) 由于 $E(s_x^2) = S_x^2$, 且根据(2.19)式, $V(s^2) = O\left(\frac{1}{n}\right)$, 因而当 n 足够大时,

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)s_x^2 \approx (n-1)S_x^2.$$

于是

$$\begin{aligned} (b - B)^2 &= \left[\frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 \\ &\approx \frac{n^2}{(n-1)^2 S_x^4} \left[\frac{1}{n} \sum_{i=1}^n \varepsilon_i (x_i - \bar{x}) \right]^2. \end{aligned}$$

根据引理 4.2 中的 3), 有

$$E(b - B)^2 = O\left(\frac{1}{n}\right).$$

这里要说明的是, 利用 Taylor 展开式可以证明当 n 大时, 用 $\frac{1}{(n-1)S_x^2}$ 近似 $\frac{1}{\sum (x_i - \bar{x})^2}$ 引起的误差的阶为 $O\left(\frac{1}{n}\right)$, 从而在其后的推导中, 这个误差可以归入高阶无穷小项中去. 有兴趣的读者可参阅 Sukhatme(1954)的书.

3) 同理, 当 n 大时, 有

$$(b - B)^4 \approx \frac{n^4}{(n-1)^4 S_x^8} \left[\frac{1}{n} \sum_{i=1}^n \varepsilon_i (x_i - \bar{x}) \right]^4.$$

从而根据引理 4.2 中的 4), 有

$$E(b - B)^4 = O\left(\frac{1}{n^2}\right).$$

$$4) \quad |E(b - B)| \leq \sqrt{E(b - B)^2} = O\left(\frac{1}{\sqrt{n}}\right). \quad \blacksquare$$

4.6.2 基本性质

定理 4.10 对简单随机抽样, 以(4.113)定义的回归估计量 \bar{y}_{lr} 有以下性质:

$$\begin{aligned} 1) \quad E(\bar{y}_{lr}) - \bar{Y} &= -\frac{1-f}{(n-1)S_y^2} \cdot \frac{\sum_{i=1}^N e_i (X_i - \bar{X})^2}{N-1} + O\left(\frac{1}{n^{3/2}}\right) \\ &= O\left(\frac{1}{n}\right), \end{aligned} \quad (4.123)$$

其中

$$e_i = Y_i - \bar{Y} - B(X_i - \bar{X}).$$

$$2) \quad \text{MSE}(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2 (1-\rho^2) + O\left(\frac{1}{n^{3/2}}\right) = O\left(\frac{1}{n}\right). \quad (4.124)$$

$$\begin{aligned} 3) \quad E(s_e^2) &\triangleq E\left\{\frac{1}{n-2} \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2\right\} \\ &= S_y^2 (1-\rho^2) + O\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \quad (4.125)$$

这个定理说明了以下事实: 当 n 大时, 有

1° \bar{y}_{lr} 是近似无偏的, 且是可用的;

2° 当 n 大时,

$$\text{MSE}(\bar{y}_{lr}) \approx V(\bar{y}_{lr}) \approx \frac{1-f}{n} S_y^2 (1-\rho^2); \quad (4.126)$$

$$\begin{aligned} 3^\circ \quad v(\bar{y}_{lr}) &= \frac{1-f}{n(n-2)} \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \\ &= \frac{1-f}{n(n-2)} \left\{ \sum_{i=1}^n (y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\} \\ &= \frac{1-f}{n} s_e^2 \end{aligned} \quad (4.127)$$

是 $V(\bar{y}_{lr})$ 或 $\text{MSE}(\bar{y}_{lr})$ 的一个近似估计, 其中 s_e^2 是样本残差方差.

定理的证明 利用上一段中的记号.

$$1) \quad E(\bar{y}_{lr}) - \bar{Y} = -E[b(\bar{x} - \bar{X})]$$

$$= -E\left[B + \frac{\sum_{i=1}^n e_i (x_i - \bar{X}) + n e (\bar{X} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] (\bar{x} - \bar{X}).$$

当 n 大时,

$$E(\bar{y}_{lr}) - \bar{Y} \approx - \frac{E\left[\sum_{i=1}^n \varepsilon_i (x_i - \bar{X})(\bar{x} - \bar{X})\right]}{(n-1)S_x^2} + \frac{nE[\bar{\varepsilon}(\bar{x} - \bar{X})^2]}{(n-1)S_x^2},$$

$$\begin{aligned} \text{而} \quad & - \frac{E\left[\sum_{i=1}^n \varepsilon_i (x_i - \bar{X})(\bar{x} - \bar{X})\right]}{(n-1)S_x^2} \triangleq \frac{nE(\bar{u} - \bar{U})(\bar{x} - \bar{X})}{(n-1)S_x^2} \\ & = \frac{1-f}{(n-1)S_x^2} \frac{\sum_{i=1}^N \varepsilon_i (X_i - \bar{X})^2}{N-1}, \\ & \left| \frac{n}{(n-1)S_x^2} E[\varepsilon(\bar{x} - \bar{X})^2] \right| \leq \frac{n}{(n-1)S_x^2} \sqrt{E(\varepsilon^2)E(\bar{x} - \bar{X})^4} \\ & = O\left(\frac{1}{n^{3/2}}\right), \end{aligned}$$

$$\text{于是} \quad E(\bar{y}_{lr}) - \bar{Y} = - \frac{1-f}{(n-1)S_x^2} \cdot \frac{\sum_{i=1}^N \varepsilon_i (X_i - \bar{X})^2}{N-1} + O\left(\frac{1}{n^{3/2}}\right).$$

正如前面已指出的那样, 由于

$$\frac{1}{\sum (x_i - \bar{x})^2} \approx \frac{1}{(n-1)S_x^2}$$

引起的误差可以归并到后面的高阶无穷小项中.

2) 令 $y'_{lr} = \bar{y} + B(\bar{X} - \bar{x})$, 则根据定理 4.8, 它是 \bar{Y} 的无偏估计, 且

$$V(y'_{lr}) = \frac{1-f}{n} S_y^2 (1-\rho^2).$$

$$\begin{aligned} \text{于是} \quad \text{MSE}(\bar{y}_{lr}) &= E(\bar{y}_{lr} - Y)^2 \\ &= E(\bar{y}_{lr} - y'_{lr} + y'_{lr} - Y)^2 \\ &= E(\bar{y}_{lr} - y'_{lr})^2 + E(y'_{lr} - Y)^2 \\ &\quad + 2E(\bar{y}_{lr} - y'_{lr})(y'_{lr} - Y) \\ &= E[(b-B)^2(\bar{X} - \bar{x})^2] + V(y'_{lr}) \\ &\quad + 2E[(y'_{lr} - Y)(b-B)(\bar{X} - \bar{x})] \\ &= O\left(\frac{1}{n^2}\right) + \frac{1-f}{n} S_y^2 (1-\rho^2) + O\left(\frac{1}{n^{3/2}}\right) \\ &= \frac{1-f}{n} S_y^2 (1-\rho^2) + O\left(\frac{1}{n^{3/2}}\right). \end{aligned}$$

这里利用了引理 4.3 中的结果.

$$3) \quad s_e^2 = \frac{1}{n-1} \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 = \frac{1}{n-1} \sum_{i=1}^n [(y_i - \bar{y}) - B(x_i - \bar{x})]^2$$

$$\text{是} \quad S_e^2 = \frac{1}{N-1} \sum_{i=1}^N [(y_i - \bar{Y}) - B(x_i - \bar{X})]^2$$

$$\begin{aligned}
 &= S_y^2 + B^2 S_x^2 - 2BS_{yx} \\
 &= S_y^2(1 - \rho^2)
 \end{aligned}$$

的无偏估计.

另一方面,

$$\begin{aligned}
 e_i - \bar{e} &= [(y_i - \bar{y}) - b(x_i - \bar{x})] + (b - B)(x_i - \bar{x}), \\
 E\left[\frac{1}{n-1} \sum_{i=1}^n (e_i - \bar{e})^2\right] &= E\left\{\frac{1}{n-1} \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2\right\} \\
 &\quad + E\left[(b - B)^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}\right] \\
 &\quad + 2E\left[(b - B) \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{n-1}\right] \\
 &\quad - 2E\left[(b - B)b \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}\right] \\
 &= E\left\{\frac{1}{n-1} \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2\right\} \\
 &\quad + O\left(\frac{1}{n}\right) + O\left(\frac{1}{\sqrt{n}}\right) + O\left(\frac{1}{\sqrt{n}}\right) \\
 &= E\left\{\frac{n-2}{n-1} s_e^2\right\} + O\left(\frac{1}{\sqrt{n}}\right).
 \end{aligned}$$

$$\begin{aligned}
 \text{从而} \quad E(s_e^2) &= E\left\{\frac{1}{n-2} \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2\right\} \\
 &= S_y^2(1 - \rho^2) + O\left(\frac{1}{\sqrt{n}}\right). \blacksquare
 \end{aligned}$$

例 4.4 (续例 4.2) 对表 4.2 所列的 $N=6$ 的人为总体, 所有可能的 15 个 $n=4$ 的简单随机样本的 y_i 对 x_i 的回归系数 b 列于表 4.3 第 7 列中, 由此可计算对总体均值 $\bar{Y}=11$ 的 15 个回归估计值 \bar{y}_{lr} (表 4.3 最后一列). 这 15 个 \bar{y}_{lr} 的均值为:

$$E(\bar{y}_{lr}) = \frac{1}{15}(10.5184 + 10.71186 + \cdots + 10.4445) = 10.9245.$$

故 y_{lr} 的实际偏倚为:

$$B(\bar{y}_{lr}) = |E(\bar{y}_{lr}) - \bar{Y}| = |10.9245 - 11| = 0.0755.$$

同样可计算 \bar{y}_{lr} 的均方误差:

$$\text{MSE}(\bar{y}_{lr}) = E(\bar{y}_{lr} - \bar{Y})^2 = 0.06776.$$

$$\text{从而} \quad V(\bar{y}_{lr}) = \text{MSE}(\bar{y}_{lr}) - [B(\bar{y}_{lr})]^2 = 0.06206.$$

与例 4.2 中的比估计 \bar{y}_R 的相应数值作比较, 对这个总体比估计的均方误差比回归估计量小得多, 这是由于前者的偏倚小, 其原因是样本量 n 太小.

4.6.3 回归估计量与简单估计量及比估计量的比较

若 n 比较大, 则根据前面的讨论, 有

$$\begin{aligned} V(\bar{y}_R) &\approx \frac{1-f}{n} S_y^2 (1 - \rho^2), \\ V(\bar{y}_R) &\approx \frac{1-f}{n} (S_y^2 + R^2 S_x^2 - 2R\rho S_y S_x), \\ V(\bar{y}) &= \frac{1-f}{n} S_y^2. \end{aligned}$$

比较上述三个式子可知:

$$V(\bar{y}_R) \leq V(\bar{y}).$$

即回归估计总是优于简单估计的, 除非 $\rho=0$. 而回归估计优于比估计的条件是:

$$\begin{aligned} -\rho^2 S_y^2 &\leq R^2 S_x^2 - 2R\rho S_y S_x \\ \Leftrightarrow (\rho S_y - R S_x)^2 &\geq 0 \Leftrightarrow (B - R)^2 \geq 0, \end{aligned} \quad (4.128)$$

因而除非 $B=R$, 否则, 回归估计优于比估计.

上面的结论只有当 n 大时才成立. 事实上, 回归估计量在小样本时的性质并不太好. 正如在例 4.4 中所述的那样, J. N. K. Rao 曾对 8 个自然总体进行 Monte-Carlo 模拟, 当 n 小时, $\text{MSE}(\bar{y}_R)/\text{MSE}(\bar{y}_R)$ 的值平均为: $n=12$, 1.15; $n=8$, 1.36; $n=6$, 1.51. 可见 n 愈小, 回归估计量的均方误差愈大. 使小样本时, 回归估计量均方误差较大的主要原因是偏倚较大, 而两者的方差相差并不显著.

例 4.5(续例 4.1) 为估计某县小麦总产量, 在全县 $N=576$ 个村中抽取 $n=24$ 个村的简单随机样本, 表 4.1 记录了样本村的小麦产量 y_i (t) 及相应的种植面积 x_i (hm²). 根据原始数据以及例 4.1 中已计算过的中间结果, 可得样本回归系数:

$$b = \frac{l_{yx}}{l_{xx}} = \frac{s_{yx}}{s_x^2} = 3.38275,$$

残差方差
$$s_e^2 = \frac{1}{n-2} (l_{yy} - b l_{yx}) = 26.428.$$

于是总产量 Y (与例 4.1 一样, 此时估计 \bar{Y} 没有什么实际意义) 的回归估

计为:

$$\begin{aligned}\hat{P}_{lr} &= N\bar{y}_{lr} = N[\bar{y} + b(\bar{X} - \bar{x})] \\ &= 576 \times [130.625 + 3.38275(37.97847 - 36.4625)] \\ &= 78193.8(\text{t}).\end{aligned}$$

\hat{P}_{lr} 的方差估计为:

$$\begin{aligned}v(\hat{P}_{lr}) &= N^2 v(\bar{y}_{lr}) = N^2 \frac{1-f}{n} s_e^2 \\ &= 576^2 \times \frac{0.95838}{24} \times 26.428 = 350117, \\ \sqrt{v(\hat{P}_{lr})} &= 591.7(\text{t}).\end{aligned}$$

例 4.1 中已计算得 $\sqrt{v(\hat{P}_x)} = 620.5(\text{t})$, $\sqrt{v(\hat{P})} = 3838.5(\text{t})$.

因此在此例中, 比估计与回归估计都比简单估计精确得多, 而回归估计与比估计的精度差别不很大, 前者稍好些.

§ 4.7 分层随机抽样中的回归估计

4.7.1 分别回归估计

同比估计情形一样, 在分层随机抽样中, 也可以考虑两种形式的回归估计. 一种是分别回归估计, 它是先在每层中对层均值或层总和作回归估计, 然后再按层权平均或相加. 具体地说, 对 \bar{Y} 的分别回归估计是:

$$\begin{aligned}\bar{y}_{lrs} &= \sum_{h=1}^L W_h \bar{y}_{lrh} \\ &= \sum_{h=1}^L W_h [\bar{y}_h + \beta_h (\bar{X}_h - \bar{x}_h)].\end{aligned}\quad (4.129)$$

而对 Y 的估计是:

$$\hat{P}_{lrs} = N\bar{y}_{lrs} = \sum_{h=1}^L N_h [\bar{y}_h + \beta_h (\bar{X}_h - \bar{x}_h)].\quad (4.130)$$

当各层的 β_h 均事先已设定时, \bar{y}_{lrs} 与 \hat{P}_{lrs} 都是无偏的, 且

$$V(\bar{y}_{lrs}) = \sum_h \frac{W_h^2(1-f_h)}{n_h} (S_{yh}^2 - 2\beta_h S_{yxh} + \beta_h^2 S_{xh}^2).\quad (4.131)$$

它在 $\beta_h = B_h = \frac{S_{yxh}}{S_{xh}^2} (h=1, 2, \dots, L)$ 时, 达到极小值:

$$\begin{aligned}V_{\min}(\bar{y}_{lrs}) &= \sum_h \frac{W_h^2(1-f_h)}{n_h} \left(S_{yh}^2 - \frac{S_{yxh}^2}{S_{xh}^2} \right) \\ &= \sum_h \frac{W_h^2(1-f_h)}{n_h} S_{yh}^2 (1-f_h^2).\end{aligned}\quad (4.132)$$

若 β_h 不能事先设定, 则取 β_h 为 B_h 的最小二乘估计:

$$\hat{b}_h = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}, \quad (4.133)$$

则当每层的 n_h 都比较大时, 有

$$V(\bar{y}_{lrh}) \approx \sum_h \frac{W_h^2(1-f_h)}{n_h} S_{yh}^2(1-\rho_h^2), \quad (4.134)$$

它可用下式估计:

$$\begin{aligned} v(\bar{y}_{lrh}) &= \sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-2)} \left[\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 - \hat{b}_h^2 \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2 \right] \\ &= \sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-2)} (n_h-1) s_{yh}^2(1-r_h^2), \end{aligned} \quad (4.135)$$

其中 r_h^2 是 h 层样本相关系数的平方.

与比估计的情形类似, 在采用分别回归估计时, 有可能由于 n_h 不够大, 而造成较大的偏倚, 使估计的均方误差较大. 此时应慎重使用.

4.7.2 联合回归估计

联合回归估计是先对 \bar{Y} 及 \bar{X} 作分层估计:

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h, \quad \bar{x}_{st} = \sum_h W_h \bar{x}_h.$$

则 \bar{Y} 与 \bar{X} 的联合回归估计分别为:

$$\hat{\bar{y}}_{lro} = \bar{y}_{st} + \beta(\bar{X} - \bar{x}_{st}), \quad (4.136)$$

$$\hat{Y}_{lro} = N \hat{\bar{y}}_{lro} = \hat{Y}_{st} + \beta(X - \hat{X}_{st}). \quad (4.137)$$

当 β 事先设定时, 它们都是无偏的, 且

$$V(\hat{\bar{y}}_{lro}) = \sum_h \frac{W_h^2(1-f_h)}{n_h} (S_{yh}^2 - 2\beta S_{yxh} + \beta^2 S_{xh}^2). \quad (4.138)$$

它在 β 取下式时达到极小值:

$$B_o = \frac{\sum_h W_h^2(1-f_h) S_{yxh} / n_h}{\sum_h W_h^2(1-f_h) S_{xh}^2 / n_h} \quad (4.139)$$

$$\triangleq \sum_h a_h B_h / \sum_h a_h, \quad (4.140)$$

其中

$$a_h \triangleq \frac{W_h^2(1-f_h)}{n_h} S_{xh}^2, \quad (4.141)$$

$$B_h = \frac{S_{yxh}}{S_{xh}^2}.$$

(4.140)式表明 B_o 是各层总体回归系数 B_h 以 a_h 为权的加权平均.

为比较分别回归估计与联合回归估计, 作最小方差差值:

$$\begin{aligned} V_{\min}(\bar{y}_{lrc}) - V_{\min}(\bar{y}_{lrs}) \\ = \sum_h \alpha_h B_h^2 - (\sum \alpha_h) B_0^2 \\ = \sum_h \alpha_h (B_h - B_0)^2 \geq 0. \end{aligned} \quad (4.142)$$

上式表明: 对最优的 B_h 与 B_0 的设定, 分别回归估计优于联合回归估计, 尤其是当各层回归系数相差大时, 分别估计效果更为显著.

当 B_0 必须从样本估计时, 我们取 B_0 的样本估计:

$$b_0 = \frac{\sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}. \quad (4.143)$$

如果样本量是按比例分配的, 又用 n_h 代替上式中的 n_h-1 , 则(4.143)式即可简化为通常的联合最小二乘估计:

$$b'_0 = \frac{\sum_h \sum_i (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h)}{\sum_h \sum_i (x_{hi} - \bar{x}_h)^2}. \quad (4.144)$$

并不是在任何情况下, b_0 与 b'_0 都是好的. 例如若 B_h 都相等, 但各层残差方差相差较大时, 则用 b_h 的与估计方差成反比的权的加权平均更为适宜.

为计算 \bar{y}_{lrc} 的方差, 注意

$$\begin{aligned} y_{lrc} - \bar{Y} &= \bar{y}_{st} - \bar{Y} + b_0(\bar{X} - \bar{x}_{st}) \\ &= [\bar{y}_{st} - \bar{Y} + B_0(\bar{X} - \bar{x}_{st})] + (b_0 - B_0)(\bar{X} - \bar{x}_{st}), \end{aligned}$$

若 b_0 的抽样误差可以忽略的话, 则

$$V(\bar{y}_{lrc}) \approx \sum_h \frac{W_h^2(1-f_h)}{n_h} (S_{y_h}^2 - 2B_0 S_{y_{xh}} + B_0^2 S_{x_h}^2), \quad (4.145)$$

它可用下式进行估计:

$$\begin{aligned} v(y_{lrc}) &\approx \sum_h \frac{W_h^2(1-f_h)}{n_h(n_h-1)} \sum_i [(y_{hi} - \bar{y}_h) - b_0(x_{hi} - \bar{x}_h)]^2 \\ &\quad \sum_h \frac{W_h^2(1-f_h)}{n_h} (s_{y_h}^2 - 2b_0 s_{y_{xh}} + b_0^2 s_{x_h}^2). \end{aligned} \quad (4.146)$$

若各层的 n_h 不太大, B_h 的变化也不大时, 宜用联合估计, 而当 B_h 的变化较大, n_h 也比较大时, 则用分别估计效果更好. 若层内回归规律性不是很强, 则除非 n_h 都相当大, 否则, 通常还是用联合估计比较保险.

4.7.3 数值例子——专业技术人员总数的调查

例 4.6 已知某市中央直属单位及市属单位 1986 年专业技术人员的总数, 欲通过抽样调查估计 1988 年年底全市专业技术人员的总数 Y . 抽样按中央直属单位与市属单位分层随机抽取. 前者抽 $n_1=15$ 个单位, 后者抽 $n_2=20$ 个单位. 数据如表 4.7 所示.

表 4.7 专业技术人员数调查 (y_i : 1988 年底的数, x_i : 1986 年底的数)

中央直属单位 $h=1$						市属单位 ($h=2$)					
i	x_i	y_i	i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	215	224	11	2158	2220	1	87	94	11	390	429
2	1082	1110	12	318	334	2	123	132	12	97	104
3	675	714	13	457	461	3	59	62	13	103	107
4	382	393	14	234	243	4	14	17	14	284	290
5	180	189	15	465	472	5	657	702	15	125	125
6	632	677				6	83	88	16	674	714
7	56	61				7	208	227	17	357	385
8	812	828				8	42	45	18	218	234
9	98	101				9	38	42	19	819	868
10	217	228				10	148	165	20	146	132

表 4.8 计算专业技术人员总数估计量所用的层内数据

	$h=1$ (中央直属单位)	$h=2$ (市属单位)	总 和
N_h	135	1228	$N=1363$
n_h	15	20	$n=35$
W_h	0.099046	0.900954	
$1-f_h$	0.059259	0.049186	
\bar{X}_h	75650	315612	$\bar{X}=391262$
\bar{X}_h	560.37037	257.01303	
\bar{y}_h	550.66667	249.60000	
\bar{x}_h	532.06667	233.60000	
s_{yh}^2	298594.238	61366.147	
s_{xh}^2	282927.781	54296.568	
s_{xyh}	290611.166	57708.411	
\hat{B}_h	1.034958	1.063493	
b_h	1.0271567	1.0623972	
r_h^2	0.999695133	0.999486648	

已知中央直属单位 $N_1 = 135$ 个, 1986 年底的总人数为 $X_1 = 75650$ 人; 市属单位 $N_2 = 1228$ 个, 1986 年底的总人数为 $X_2 = 315612$ 人。

我们对上述数据按分别比估计、联合比估计、分别回归估计与联合回归估计以及差估计等方法对该市 1988 年专业技术人员总数 Y 作出估计, 同时计算各估计量的精度。为此, 先就中央直属单位及市属单位两层的样本数据计算有关层的中间结果如表 4.8 (包括某些已知量) 所示。

一、分别比估计

$$\begin{aligned}\hat{Y}_{Rz} &= \sum_h \frac{y_h}{x_h} X_h = \sum_h \hat{R}_h X_h \\ &= 1.034958 \times 75650 + 1.068493 \times 315612 = 415524, \\ v(\hat{Y}_{Rz}) &= \sum_h \frac{N_h^2(1-f_h)}{n_h} (s_{yh}^2 + \hat{R}_h^2 s_{xh}^2 - 2\hat{R}_h s_{yxh}) \\ &= 116910.2 + 2465413.8 = 2582324, \\ \sqrt{v(\hat{Y}_{Rz})} &= 1607.\end{aligned}$$

二、联合比估计

$$\begin{aligned}\hat{Y}_{st} &= \sum_h N_h \bar{y}_h = 380848.8, \\ \hat{X}_{st} &= \sum_h N_h \bar{x}_h = 358689.8, \\ \hat{R}_o &= \frac{\hat{Y}_{st}}{\hat{X}_{st}} = 1.0617776, \\ \hat{Y}_{Rc} &= \frac{\hat{Y}_{st}}{\hat{X}_{st}} X = 1.0617776 \times 391262 = 415433, \\ v(\hat{Y}_{Rc}) &= \sum_h \frac{N_h^2(1-f_h)}{n_h} (s_{yh}^2 + \hat{R}_o^2 s_{xh}^2 - 2\hat{R}_o s_{yxh}) \\ &= 464561.06 + 2341109.41 = 2805670, \\ \sqrt{v(\hat{Y}_{Rc})} &= 1675.\end{aligned}$$

三、分别回归估计 (当 β_h 采用样本回归系数 b_h 时)

$$\begin{aligned}\hat{Y}_{trs} &= \sum_h N_h \bar{y}_{trh} = \sum_h N_h [y_h + b_h (\bar{X}_h - \bar{x}_h)] \\ &= 78264.8 + 337066.6 = 415331, \\ v(\hat{Y}_{trs}) &= \sum_h \frac{N_h^2(1-f_h)}{n_h} s_y^2 \\ &= \sum_h \frac{N_h^2(1-f_h)}{n_h} \cdot \frac{n_h-1}{n_h} \cdot s_{yh}^2 (1-r_h^2) \\ &= 105876.3 + 2166398.6 = 2572275, \\ \sqrt{v(\hat{Y}_{trs})} &= 1604.\end{aligned}$$

四、联合回归估计(当 β 采用样本数据估计时)

$$b_c = \frac{\sum_h \frac{W_h^2(1-f_h)}{n_h} s_{yxh}}{\sum_h \frac{W_h^2(1-f_h)}{n_h} s_{xh}^2} = \frac{2472.9607}{2332.2757} = 1.0603209,$$

$$\begin{aligned}\hat{Y}_{lro} &= \hat{Y}_{st} + b_c(X - \hat{X}_{st}) \\ &= 380848.8 + 1.0603209(391262 - 358689.8) \\ &= 415386,\end{aligned}$$

$$\begin{aligned}v(\hat{Y}_{lro}) &= \sum_h \frac{N_h^2(1-f_h)}{n_h} (s_{yh}^2 - 2b_c s_{yxh} + b_c^2 s_{xh}^2) \\ &= 434389 - 2362087 = 2796476, \\ \sqrt{v(\hat{Y}_{lro})} &= 1672.\end{aligned}$$

五、差估计

由于回归系数接近 1, 故可用差估计, 也即 β (或 β_h) 设定为常数 1 的回归估计. 注意此时分别估计与联合估计结果相同.

$$\hat{Y}_d = \sum_h [X_h + N_h(\bar{y}_h - x_h)] = \hat{Y}_{st} + X - \hat{X}_{st} = 413421,$$

$$\begin{aligned}v(\hat{Y}_d) &= \sum_h \frac{N_h^2(1-f_h)}{n_h} (s_{yh}^2 + s_{xh}^2 - 2s_{yxh}) \\ &= 323661 + 18238302 = 18561963, \\ \sqrt{v(\hat{Y}_d)} &= 4308.\end{aligned}$$

为了对以上五种估计及其精度作比较, 先将 Y 的估计值及相应的标准差的估计值列于表 4.9.

表 4.9 专业技术人员总数各种估计值的比较

估计方法	估计值 \hat{Y}	\hat{Y} 的标准差估计
分别比估计	415524	1607
联合比估计	415433	1675
分别回归估计	415331	1604
联合回归估计	415386	1672
差估计	413421	4308

从表面上的数值看, 此例的比估计与回归估计效果一致, 这是因为 \hat{R}_0 与 b_0 (\hat{R}_h 与 b_h) 数值相当接近的缘故. 其中分别估计比联合估计效果稍好, 这是因为 \hat{R}_1 与 \hat{R}_2 (b_1 与 b_2) 的数值仍有一定的差别. 不过, 我们注意到比估计与回归估计(当回归系数用样本值时)都是有偏的, 而且这里的样本量 n_h 都不大, 因此用比估计特别是联合比估计更保险些. 基于同样的理由, 差估计尽管标准差较大, 但由于它是无偏的, 因此仍有采用

的价值.

§ 4.8 多变量比估计与回归估计

从前面几节的讨论中看到, 当有辅助变量可资利用时, 比估计与回归估计比简单估计能较大程度地提高估计精度. 当可供使用的辅助变量不止一个时, 上面的结果能容易地推广到多变量的情形, 本节主要介绍多变量比估计与回归估计的思想及方法, 它们的性质与单变量的情形十分类似. 前几节的许多结果可直接照搬过来, 故在本节中就不详细叙述了.

4.8.1 多变量比估计

Olkin (1958) 首先将比估计推广到有 p 个辅助变量 $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_p$ 的情形.

定义 4.6 设 \bar{y}_{Rk} 是 \bar{Y} 对第 k 个辅助变量 \mathcal{X}_k 的比估计, w_k 是适当选取的权, $\sum_{k=1}^p w_k = 1$, 则 \bar{Y} 的多变量比估计为:

$$\bar{y}_{MR} = \sum_{k=1}^p w_k \bar{y}_{Rk} = \sum_{k=1}^p w_k \frac{\bar{y}}{\bar{x}_k} \bar{X}_k, \quad (4.117)$$

其中 $\bar{y}, \bar{x}_1, \dots, \bar{x}_p$ 是相应变量的样本均值, $\bar{X}_1, \dots, \bar{X}_p$ 是辅助变量总体的均值.

这里的主要问题是关于权 w_k 的选取. 确定 w_k 的原则是使 $V(\bar{y}_{MR})$ 达到极小. 例如在 $p=2$ 的情形

$$\begin{aligned} V(\bar{y}_{MR}) &= w_1^2 V(\bar{y}_{R1}) + 2w_1 w_2 \text{Cov}(\bar{y}_{R1}, \bar{y}_{R2}) + w_2^2 V(\bar{y}_{R2}) \\ &\triangleq w_1^2 V_{11} + 2w_1 w_2 V_{12} + w_2^2 V_{22}. \end{aligned} \quad (4.148)$$

在约束条件 $w_1 + w_2 = 1$ 之下极小化上式, 用 Lagrange 乘子法即是极小化:

$$V^* = w_1^2 V_{11} + 2w_1 w_2 V_{12} + w_2^2 V_{22} + 2\lambda(w_1 + w_2 - 1), \quad (4.149)$$

$$\frac{\partial V^*}{\partial w_1} = 2w_1 V_{11} + 2w_2 V_{12} + 2\lambda = 0,$$

$$\frac{\partial V^*}{\partial w_2} = 2w_1 V_{12} + 2w_2 V_{22} + 2\lambda = 0.$$

将上两式相减, 得:

$$w_1(V_{11} - V_{12}) + w_2(V_{12} - V_{22}) = 0,$$

又将 $w_2 = 1 - w_1$ 代入, 有

$$w_1(V_{11} - V_{12}) + V_{12} - V_{22} - w_1V_{12} + w_2V_{22} = 0,$$

从而可解得

$$w_1 = \frac{V_{22} - V_{12}}{V_{11} + V_{22}} \frac{1}{2V_{12}}, \quad w_2 = \frac{V_{11} - V_{12}}{V_{11} + V_{22}} \frac{1}{2V_{12}}. \quad (4.150)$$

此时, 最小方差为

$$V_{\min}(\bar{y}_{MR}) = \frac{V_{11}V_{22}}{V_{11} + V_{22}} \frac{V_{12}^2}{2V_{12}}. \quad (4.151)$$

在实际问题中, V_{11} 、 V_{12} 与 V_{22} 都用相应的样本估计量代替. 根据 § 4.1 中的结果, 可取

$$v_{11} = \frac{1-f}{n} \hat{Y}^2 (c_y^2 + c_1^2 - 2c_{y1}),$$

$$v_{22} = \frac{1-f}{n} \hat{Y}^2 (c_y^2 + c_2^2 - 2c_{y2}),$$

$$v_{12} = \frac{1-f}{n} \hat{Y}^2 (c_y^2 + c_{12} - c_{y1} - c_{y2}).$$

其中

$$c_y^2 = \frac{s_y^2}{\bar{y}^2}, \quad c_1^2 = \frac{s_{x_1}^2}{\bar{x}_1^2}, \quad c_2^2 = \frac{s_{x_2}^2}{\bar{x}_2^2},$$

$$c_{12} = \frac{s_{x_1}s_{x_2}}{\bar{x}_1\bar{x}_2}, \quad c_{y1} = \frac{s_{yx_1}}{\bar{y}\bar{x}_1}, \quad c_{y2} = \frac{s_{yx_2}}{\bar{y}\bar{x}_2}.$$

而 \hat{Y} 是 \bar{Y} 的适当估计, 例如用相应的单变量比估计 \bar{y}_{R1} 或 \bar{y}_{R2} . 注意到根据 (4.150) 式, 在确定 w_1 、 w_2 时常数因子 $\frac{1-f}{n} \hat{Y}^2$ 不起作用, 而在计算所得的 \bar{y}_{MR} 的方差时, \hat{Y} 则可用 \bar{y}_{MR} 值代替, 详见后面 4.8.3 段中的数值例子.

对一般的 p , 仍令

$$V_{kl} = \begin{cases} V(\bar{y}_{Rk}), & k = l, \\ \text{COV}(\bar{y}_{Rk}, \bar{y}_{Rl}), & k \neq l. \end{cases} \quad (4.152)$$

记

$$(V^{kl}) = (V_{kl})^{-1}, \quad (4.153)$$

则

$$w_k = \frac{\sum_{l=1}^p V^{kl}}{\sum_{k=1}^p \sum_{l=1}^p V^{kl}} \quad (k=1, 2, \dots, p), \quad (4.154)$$

$$V_{\min}(\bar{y}_{MR}) = \left[\sum_{k=1}^p \sum_{l=1}^p V^{kl} \right]^{-1}. \quad (4.155)$$

4.8.2 多变量回归估计

将一个辅助变量情形的回归估计推广到多个辅助变量情形时, 有两种方法, 一种是与 Olkin 的多变量比估计相类似的, 采用加权法, 这就是由 Des Raj (1965) 最早提出的.

定义 4.7 若 \bar{y}_{lrk} 是 \bar{Y} 对第 k 个辅助变量 \mathcal{X}_k 的回归估计, w_k 是适当选取的权, $\sum_{k=1}^p w_k = 1$, 则 \bar{Y} 的多变量回归估计为:

$$\bar{y}_{MLR} = \sum_{k=1}^p w_k \bar{y}_{lrk} = \sum_{k=1}^p w_k [\bar{y} + \beta_k (\bar{X}_k - \bar{x}_k)]. \quad (4.156)$$

式中 β_k 可以事先设定, 也可取样本回归系数. 确定 w_k 的原则与方法也与 Olkin 的多变量比估计类似.

另一个更为直观的方法是早在 1947 年由 B. Ghosh 提出的利用 \mathcal{Y} 对 $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_p$ 的多元线性回归, 他提出的估计量形式如下:

定义 4.8 \bar{Y} 的多元(线性)回归估计为:

$$\bar{y}_{MLR} = \bar{y} + \sum_{k=1}^p \beta_k (\bar{X}_k - \bar{x}_k). \quad (4.157)$$

注意: 当 β_k 都是事先设定时, 两种形式的估计量一致. 因此通常是取 β_k 为 \mathcal{Y} 对 \mathcal{X}_k 的样本(偏)回归系数 $b_k (k=1, 2, \dots, p)$, 即

$$\bar{y}_{MLR} = \bar{y} + \sum_{k=1}^p b_k (\bar{X}_k - \bar{x}_k) = \bar{y} - \sum_{k=1}^p b_k (\bar{x}_k - \bar{X}_k), \quad (4.158)$$

\bar{y}_{MLR} 是渐近无偏的, 偏倚的阶为 $O\left(\frac{1}{n}\right)$, 且

$$\text{MSE}(\bar{y}_{MLR}) = \frac{1-f}{n} S_y^2 (1-\rho^2) + O\left(\frac{1}{n^{3/2}}\right). \quad (4.159)$$

其中 ρ^2 是 \mathcal{Y} 对 $\mathcal{X}_1, \dots, \mathcal{X}_p$ 的复相关系数的平方. 因而当 n 大时, $\text{MSE}(\bar{y}_{MLR})$ 或 $V(\bar{y}_{MLR})$ 可用下式估计:

$$v(\bar{y}_{MLR}) = \frac{1-f}{n} s_e^2, \quad (4.160)$$

式中 s_e^2 是残差方差:

$$s_e^2 = \frac{1}{n-p-1} \left[\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{k=1}^p b_k \sum_{i=1}^n (y_i - \bar{y})(x_{ik} - \bar{x}_k) \right]. \quad (4.161)$$

4.8.3 数值例子——农作物估产调查

例 4.7 为更精确地估计某地区皮棉的总产量, 在对总体包含

的 $N = 301$ 个村庄中按简单随机抽样抽取 $n = 18$ 个村庄, 在记录这些村中皮棉的实际产量 y_i 的同时, 记录皮棉的播种面积 x_{1i} 及所采用的良种比例 x_{2i} . 已知该地区的皮棉种植总面积 $X_1 = 74500(\text{hm}^2)$, 而采用良种的平均比例 $\bar{X}_2 = 40.10(\%)$, 样本数据见表 4.10.

表 4.10 18 个村庄的皮棉产量及其播种面积与良种比例

村庄号 i	产量 y_i (t)	播种面积 x_{1i} (hm^2)	良种比例 x_{2i} (%)
1	12.00	24.0	30
2	11.88	26.4	30
3	12.50	25.0	32
4	14.70	21.0	55
5	10.00	12.5	58
6	10.80	14.0	50
7	15.02	35.0	35
8	22.00	44.0	36
9	21.40	26.0	63
10	5.46	16.0	18
11	7.60	20.0	20
12	10.66	20.0	38
13	21.60	35.0	42
14	8.41	18.0	33
15	27.02	38.0	53
16	17.00	25.0	43
17	13.64	23.0	36
18	6.65	17.0	20

一、基本数据及中间结果

$N = 301$	$n = 18$	$\frac{1-f}{n} = 0.0522382$
$\sum y_i = 248.34$	$\sum x_{1i} = 489.9$	$\sum x_{2i} = 692$
$\bar{y} = 13.7967$	$\bar{x}_1 = 2.443889$	$\bar{x}_2 = 38.4444$
$l_{yy} = 603.258$	$l_{11} = 1269.54$	$l_{22} = 2974.44$
$s_y^2 = 35.4858$	$s_1^2 = 74.678987$	$s_2^2 = 174.9671$
$l_{y1} = 718.4483$	$l_{y2} = 790.700$	$l_{12} = 196.2890$
$s_{y1} = 42.26167$	$s_{y2} = 46.5118$	$s_{12} = 11.54641$
$r_{y1} = 0.82096$	$r_{y2} = 0.590278$	$r_{12} = 0.101011$

这里的 $l_{..}$ 表示离差平方和, $r_{..}$ 表示(单)相关系数. 由于篇幅限制, 各数值的有效数字没有列出更多, 而以下的实际计算是根据更多的有效

数字计算的.

二、比估计

$$\hat{P}_{R1} = \frac{\bar{y}}{\bar{x}_1} X_1 = 4205.80,$$

$$\hat{P}_{R2} = \frac{\bar{y}}{\bar{x}_2} X_2 = \frac{\bar{y}}{\bar{x}_2} N \bar{X}_2 = 4331.63,$$

$$c_y^2 = \frac{s_y^2}{\bar{y}^2} = 0.186425, \quad c_{y1} = \frac{s_{y1}}{\bar{y}\bar{x}_1} = 0.125340,$$

$$c_1^2 = \frac{s_1^2}{\bar{x}_1^2} = 0.125036, \quad c_{y2} = \frac{s_{y2}}{\bar{y}\bar{x}_2} = 0.087691,$$

$$c_2^2 = \frac{s_2^2}{\bar{x}_2^2} = 0.118983, \quad c_{12} = \frac{s_{12}}{\bar{x}_1\bar{x}_2} = 0.012289,$$

$$u_{11} \triangleq c_y^2 + c_1^2 - 2c_{y1} = 0.060781,$$

$$u_{22} \triangleq c_y^2 + c_2^2 - 2c_{y2} = 0.129426,$$

$$u_{12} \triangleq c_y^2 + c_{12} - c_{y1} - c_{y2} = 0.014917,$$

$$w_1 = \frac{V_{22} \cdot V_{12}}{V_{11} + V_{22} - 2V_{12}} = \frac{u_{22} - u_{12}}{u_{11} - u_{22} - 2u_{12}} = 0.6568,$$

$$w_2 = \frac{V_{11} \cdot V_{12}}{V_{11} + V_{22} - 2V_{12}} = \frac{u_{11} - u_{12}}{u_{11} + u_{22} - 2u_{12}} = 0.3432.$$

从而

$$\hat{P}_{MR} = w_1 \hat{P}_{R1} + w_2 \hat{P}_{R2} = 4248.98,$$

$$\begin{aligned} v(\hat{P}_{MR}) &= \frac{V_{11}V_{22} - V_{12}^2}{V_{11} + V_{22} - 2V_{12}} \\ &= \frac{1-f}{n} [\hat{P}_{MR}]^2 \frac{u_{11}u_{22} - u_{12}^2}{u_{11} + u_{22} - 2u_{12}} = 33014.94, \end{aligned}$$

$$\sqrt{v(\hat{P}_{MR})} = 181.70,$$

$$v(\hat{P}_{R1}) = v_{11} = \frac{1-f}{n} (\hat{P}_{R1})^2 u_{11} = 56158.11,$$

$$\sqrt{v(\hat{P}_{R1})} = 236.98,$$

$$v(\hat{P}_{R2}) = v_{22} = \frac{1-f}{n} (\hat{P}_{R2})^2 u_{22} = 126844.51,$$

$$\sqrt{v(\hat{P}_{R2})} = 356.15.$$

三、回归估计

单变量回归估计

$$b_1 = \frac{l_{y1}}{l_{11}} = 0.5659,$$

$$\bar{y}_{lr1} = \bar{y} + b_1(\bar{X}_1 - \bar{x}_1) = 13.9732,$$

$$\hat{P}_{lr1} = N\bar{y}_{lr1} = 4205.93,$$

$$v(\hat{P}_{lr1}) = \frac{N^2(1-f)}{n} s_e^2 = \frac{N^2(1-f)}{n(n-2)} \left(l_{yy} - \frac{l_{y1}^2}{l_{11}} \right) = 58169,$$

$$\sqrt{v(\hat{P}_{lr1})} = 241.18,$$

$$b_2 = \frac{l_{y2}}{l_{22}} = 0.26583,$$

$$\bar{y}_{lr2} = \bar{y} + b_2(\bar{X}_2 - \bar{x}_2) = 14.2368,$$

$$\hat{P}_{lr2} = N\bar{y}_{lr2} = 4285.28,$$

$$v(\hat{P}_{lr2}) = \frac{N^2(1-f)}{n(n-2)} \left(l_{yy} - \frac{l_{y2}^2}{l_{22}} \right) = 116345,$$

$$\sqrt{v(\hat{P}_{lr2})} = 341.09.$$

二元回归估计: 决定二元回归方程的回归系数 b_1 、 b_2 的正规方程为:

$$\begin{cases} l_{11}b_1 + l_{12}b_2 = l_{1y}, \\ l_{12}b_1 + l_{22}b_2 = l_{2y}. \end{cases}$$

由此可解出

$$b_1 = 0.5302, \quad b_2 = 0.23084.$$

经方差分析, 得到二元回归的残差方差:

$$s_e^2 = \frac{1}{n-3} [l_{yy} - b_1l_{y1} - b_2l_{y2}] = 2.6541,$$

$$\hat{P}_{MLR} = N[y + b_1(\bar{X}_1 - \bar{x}_1) + b_2(\bar{X}_2 - \bar{x}_2)] = 4317.63,$$

$$v(\hat{P}_{MLR}) = \frac{N^2(1-f)}{n} s_e^2 = 12560.4,$$

$$\sqrt{v(\hat{P}_{MLR})} = 112.07.$$

四、结果的比较

上述结果可列成表 4.11.

表 4.11 两变量比估计和回归估计与单变量相应估计的比较

估计方法	估计量 $\hat{Y}(t)$	\hat{Y} 的标准差估计 t
对 \mathcal{X}_1 的比估计	4205.80	236.98
对 \mathcal{X}_2 的比估计	4331.63	356.15
对 \mathcal{X}_1 的回归估计	4205.93	241.18
对 \mathcal{X}_2 的回归估计	4285.28	341.09
两变量比估计	4248.98	181.70
两变量回归估计	4317.63	112.07

在本例中两变量的比估计与回归估计在精度上有较大程度的提高.

§ 4.9 二相抽样中的比估计与回归估计

在对总体均值 Y (或总和 Y) 的比估计与回归估计中都需要已知辅助变量的均值 X (或总和 X). 若 X 未知, 则正如在分层抽样中层权未知的情形一样, 可以用二相抽样. 此时第一相样本用于估计 X , 而从第一相样本中随机抽出的第二相(子)样本则用来构造通常意义的比估计和回归估计.

4.9.1 比估计

从总体中抽取一个样本量为 n' 的简单随机样本, 仅对辅助变量加以测定, 获得样本均值 \bar{x}' , 以此作为 X 的估计. 在这个第一相样本中再抽取一个样本量为 n 的简单随机样本 (抽样比 $f = \frac{n}{n'}$ 事先确定), 获得样本均值 \bar{y} 与 \bar{x} , 则二相样本中的比估计定义为:

$$\bar{y}_R = \frac{\bar{y}}{\bar{x}'} \bar{x}' \triangleq \hat{R}x'. \quad (4.162)$$

记第一相样本中 y_i 的均值为 \bar{y}' , $R' = \frac{\bar{y}'}{\bar{x}'}$, 它们都是未知的. 根据两步抽样均值与方差的一般公式 (§ 3.8):

$$E(\bar{y}_R) = E_1 E_2(\bar{y}_R) = E_1[\bar{x}' F_2(\bar{y}_R)],$$

当 n 足够大时, $E_2(\hat{R}) \approx R' = \frac{\bar{y}'}{\bar{x}'}$, 从而

$$E(\bar{y}_R) \approx E_1(\bar{y}') = Y. \quad (4.163)$$

为求 \bar{y}_R 的近似方差, 注意到 (当 n 足够大时)

$$V_2(\bar{y}_R) \approx \frac{1}{n} f s_{g'}^2 = \left(\frac{1}{n} - \frac{1}{n'} \right) s_{g'}^2.$$

其中 $s_{g'}^2$ 是 $g_i = y_i - R'x_i$ 的 (第一相) 样本方差, 它的均值 (如果忽略 R' 与 $R = Y/X$ 的差异) 是 $G_1 = Y - RX$ 的总体方差 $S_g^2 = S_y^2 + R^2 S_x^2 - 2RS_{yx}$. 于是

$$\begin{aligned} V(\bar{y}_R) &= V_1 E_2(\bar{y}_R) + E_1 V_2(\bar{y}_R) \\ &\approx V_1(\bar{y}') + \left(\frac{1}{n} - \frac{1}{n'} \right) E_1(s_{g'}^2) \end{aligned}$$

$$\begin{aligned}
&= \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) (S_y^2 + R^2 S_x^2 - 2RS_{yx}) \\
&\approx \frac{s_y^2}{n} + \left(\frac{1}{n} - \frac{1}{n'} \right) (R^2 s_x^2 - 2Rs_{yx}).
\end{aligned} \tag{4.164}$$

$V(\bar{y}_R)$ 可用下式估计.

$$v(\bar{y}_R) = \frac{s_y^2}{n} + \left(\frac{1}{n} - \frac{1}{n'} \right) (\hat{R}^2 s_x^2 - 2\hat{R}s_{yx}). \tag{4.165}$$

4.9.2 回 归 估 计

与比估计的情形一样, 样本量为 n' 的第一相样本仅测 x'_i 获得 \bar{X} 的估计 \bar{x}' , 从第二相样本中求得 \bar{y} 、 \bar{x} 及样本回归系数 b , 于是 \bar{Y} 的回归估计定义为

$$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x}), \tag{4.166}$$

$$E_2(\bar{y}_{lr}) = E_2(\bar{y}) + E_2[b(\bar{x}' - \bar{x})] \approx \bar{y}',$$

于是

$$E(\bar{y}_{lr}) = E_1 E_2(\bar{y}_{lr}) \approx E_1(\bar{y}') = \bar{Y}, \tag{4.167}$$

$$V_2(\bar{y}_{lr}) \approx \frac{1-f}{n} s_{e'}^2 = \left(\frac{1}{n} - \frac{1}{n'} \right) s_{e'}^2.$$

其中 $s_{e'}^2 = (1-\rho'^2)s_{y'}^2$ 是第一相样本残差方差, 它的均值近似等于总体残差方差 $S_{e'}^2 = (1-\rho^2)S_y^2$. 于是

$$\begin{aligned}
V(\bar{y}_{lr}) &= V_1 E_2(\bar{y}_{lr}) + E_1 V_2(\bar{y}_{lr}) \\
&\approx V_1(\bar{y}') + \left(\frac{1}{n} - \frac{1}{n'} \right) E(s_{e'}^2) \\
&\approx \left(\frac{1}{n'} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'} \right) S_y^2 (1-\rho^2) \\
&\approx \frac{S_y^2}{n} - \left(\frac{1}{n} - \frac{1}{n'} \right) \rho^2 S_y^2.
\end{aligned} \tag{4.168}$$

它可用下式进行估计.

$$v(\bar{y}_{lr}) = \frac{s_y^2}{n} - \left(\frac{1}{n} - \frac{1}{n'} \right) r^2 s_y^2. \tag{4.169}$$

第5章

不等概率抽样

§ 5.1 一般描述

5.1.1 不等概率抽样的必要性

前几章讨论的简单随机抽样与分层随机抽样有一个共同的特点：总体(或层)中的每个单元入样的概率都相等。如果总体中的每个单元在该总体中的地位(或重要性)相差不多，则这种基于等概率的抽样是理所当然的选择。等概率抽样不仅实施简单，而且相应的数据处理公式也简单。但是在许多实际问题中，我们还需要使用不等概率抽样(sampling with unequal probabilities)。一种情况是调查的总体单元与抽样总体的单元可能不一致。例如，某学校欲对学生的家庭情况进行调查，调查总体是全校学生的家庭。在这些家庭中，许多家庭只有一个孩子在该校就读，但也有些家庭有两个或两个以上的孩子在该校就读。从抽样角度来说，将学生作为抽样单元是方便的，因为相应的抽样框是现成的。而另一方面，从调查角度而言，对每个(学生)家庭实行等概率抽样又是合理的。这样就产生了一个问题：若对学生实行等概率抽样，则每个家庭被抽中的概率并不相等。例如有两个孩子在该学校就读的家庭入样的概率是只有一个孩子在该校就读的家庭入样概率的两倍。因此，为了使每个家庭入样的概率相等，就只能对学生进行不等概率抽样。方法是：对每个学生登记其家庭在该校就读的学生人数，每个学生的家庭入样的概率应与这个数字成反比。

另一种需要用到不等概率抽样的情况是，抽样单元在总体中所占的地位不一致。例如若用抽样方法估计全国科技人员在近五年内的流动总数，那么大的单位(研究所、高等院校、企业单位等)显然比小单位重要得多。类似的例子还有通过对企业的调查估计某地区某一时期的总产值，对商业网点调查估计该地区的商品零售总额等等。在这些例子中，对单位(包括工厂、商店)进行等概率抽样，估计效果一般不会很好。若对单位进行不等概率抽样，使大单位入样的概率大，小单位入样的概率小，就可

大大提高估计的精度。单位的大小可用适当的量来表示,例如研究所的科技人员数,企业与商店的固定资产或流动资金总额等。最重要的一种不等概率抽样乃是使每个单元入样的概率与该单元的大小成比例的抽样(sampling with probabilities proportional to sizes)。

第三种需用不等概率抽样的情况是为了改善估计量的特性。在 § 4.4 中提到的 Lahiri 比估计量即是其中一个例子:每个可能的样本若被抽中的概率与样本中单元的辅助变量之和成正比的话,则按此进行不等概率抽样所得到的样本,用通常的比估计方法所得的估计量是无偏的。

总之,在实际工作中需要我们经常采用不等概率抽样。在以后几章中可以看到对于整群抽样、多阶抽样及系统抽样,不等概率抽样是一种相当常用的抽样方法。另外,从上面列举的情况也可看到,凡需使用不等概率抽样的场合,必须提供总体单元的某种辅助信息,例如每个单元的“大小”度量 M_i 或辅助变量 X_i 等。

5.1.2 不等概率抽样的分类

不等概率抽样可按多种原则进行分类。鉴于不等概率抽样同时会带来目标量估计及其方差估计的复杂性,为了简化起见,人们常使用放回抽样:每次在总体(或层)中按一定概率抽取一个单元,抽取后放回总体,再进行下一次抽样,每次抽样都是独立的。在另外一些场合,为使抽样的效率更高,也使用多种不放回抽样。其代价是:由于丧失了独立性,无论是抽样方法还是方差估计,都要比放回抽样繁复得多。另一种分类是:视每次抽样(放回抽样的情形)概率或每个单元的入样概率(不放回抽样的情形)是否严格地与单元的大小成比例。另外,看样本量 n 是固定的还是随机的。最重要的情形乃是当 n 固定,且上述概率与单元大小严格成比例的不等概率抽样。以后我们将这种情形的放回抽样称为 FPS 抽样,称相应的不放回抽样为 π PS 抽样。

对于不放回抽样,又有以下几种抽取方式:

一、逐个抽取方法(draw-by-draw procedure)

每次从尚未入样的单元中以一定的概率抽取一个单元。这个概率通常与已经入样的单元有关,若无关,则称这组概率为工作概率(working probabilities)。

二、重抽方法(rejective procedure)

以一定概率逐个进行放回抽样,若一旦抽到重复单元,则放弃所有已

抽到的单元,重新抽取,直至抽到规定数目且所有入样单元都不同为止。

三、全样本方法(whole sample procedure)

对每个可能样本规定一个被抽中的概率,按这个概率一次抽取整个样本。

四、系统抽取方法(systematic procedure)

将总体单元按某种顺序排列,并将规定的单元入样概率(或其倍数)累计起来,并确定抽样间隔,在这个范围内产生一个随机数以确定初始单元,然后按上述抽样间隔确定其余的样本单元。

除了上述以外,还有其他一些抽样方法。但在本章中,我们只介绍若干常用且较为典型的方法,其中系统抽取方法将在第 8 章中再作介绍。

§ 5.2 放回不等概率抽样与 Hansen-Hurwitz 估计量

5.2.1 多项抽样、PPS 抽样及其实施方法

定义 5.1 设 Z_1, Z_2, \dots, Z_N 是一组概率, $\sum_{i=1}^N Z_i = 1$, 按这组概率对总体中的 N 个单元进行放回抽样,每次抽到第 i 个单元的概率为 Z_i , 独立地进行这样的抽样 n 次,则称这种不等概率抽样为多项抽样(multinomial sampling)。

上述抽样所以称为“多项抽样”是因为若令 t_i 是总体中第 i 个单元在 n 次抽样中被抽中的次数,则 (t_1, t_2, \dots, t_N) 的联合分布是以下的多项分布(为简化记号起见,仍以 t_i 记它的实际取值):

$$\frac{n!}{t_1! t_2! \cdots t_N!} Z_1^{t_1} Z_2^{t_2} \cdots Z_N^{t_N}, \quad \sum_{i=1}^N t_i = n. \quad (5.1)$$

根据多项分布的性质(证明完全与引理 2.3 类似,这里从略),有

$$E(t_i) = nZ_i, \quad (5.2)$$

$$V(t_i) = nZ_i(1 - Z_i), \quad (5.3)$$

$$\text{Cov}(t_i, t_j) = -nZ_i Z_j \quad (i \neq j). \quad (5.4)$$

特别当每个单元具有一个说明其大小或规模(size)的度量 M_i 时,例如单位的职工人数、农场的耕地面积、工厂的产值或该单元调查指标在上一次普查时的数值等等,则可取

$$Z_i = \frac{M_i}{M_0}, \quad (5.5)$$

其中 $M_0 = \sum_{i=1}^N M_i$ 是总体中所有单元的“大小”之和, 显然, 此时有 $\sum_{i=1}^N Z_i = 1$. 这时的多项抽样由于每个单元在每次抽样中的入样概率与单元大小成比例, 故称为(放回的)与大小成比例的概率抽样(sampling with probability proportional to size), 也即在前节提到过的 PPS 抽样.

多项抽样是最简单的不等概率抽样, 最早由 Hansen 与 Hurwitz (1943) 提出, 但“多项抽样”这个名称则迟至 1962 年由 Hartley 与 Rao 提出.

实施多项抽样有两种方法: 一是所谓代码法. 在 PPS 抽样情形, 通过对 M_i 的累计, 对每个单元赋以一个与 M_i 相等的代码数(假定所有的 M_i 都为整数, 若不然, 可乘以某个倍数. 对一般的多项抽样, 也总可找到这样一个整数 M_0 , 使所有的 $M_0 Z_i$ 皆为整数). 每次抽样产生一个 $[1, M_0]$ 之间的随机数字(整数), 设为 m , 则代码 m 所在的单元入样.

例 5.1 设某个总体共有 $N=8$ 个单元, 相应的大小 M_i 及赋予的代码如表 5.1 所示.

表 5.1 用代码法进行多项 (PPS) 抽样

i	M_i	累计 M_i	代码
1	5	5	1~5
2	4	9	6~9
3	6	15	10~15
4	4	19	16~19
5	2	21	20~21
6	1	22	22
7	3	25	23~25
■	7	32	26~32
Σ	$M_0=32$		

若 $n=2$, 则先在 $[1, 32]$ 中产生一个随机数, 设为 17, 于是第 4 个单元入样; 再在 $[1, 32]$ 中产生第二个随机数, 设为 25, 则第 7 个单元入样. 代码法对 N 不太大时是适用的, 但当 M 很大时, 就很不方便. 此时可用 Lahiri 提出的方法. 方法如下, 令 $M^* = \max_{1 \leq i \leq N} \{M_i\}$, 每次抽取一个 $[1, N]$ 范围内的随机数 i 及 $[1, M^*]$ 范围内的随机数 m , 若 $M_i \geq m$, 则第 i 个单元入样; 否则, 重抽 (i, m) . 此时第 i 个单元的入样的概率与 M_i 成正比, 从而 $Z_i = M_i / M_0$.

5.2.2 Hansen-Hurwitz 估计量及其性质

定理 5.1 若 y_1, y_2, \dots, y_n 是按 Z_i 为入样概率的多项抽样抽得的样本指标值, 相应的 Z_i 值为 z_1, z_2, \dots, z_n , 则总体总和 Y 的以下估计 (称为 Hansen-Hurwitz 估计)

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i} \quad (5.6)$$

是无偏的, 且

$$V(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 \quad (5.7)$$

$$= \frac{1}{n} \left(\sum_{i=1}^N \frac{Y_i^2}{Z_i} - Y^2 \right) \quad (5.8)$$

$$= \frac{1}{n} \sum_{i=1}^N \sum_{j=1}^N Z_i Z_j \left(\frac{Y_i}{Z_i} - \frac{Y_j}{Z_j} \right)^2. \quad (5.9)$$

又若 $n > 1$, 则

$$v(\hat{Y}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{HH} \right)^2 \quad (5.10)$$

是 $V(\hat{Y}_{HH})$ 的无偏估计.

证明 1 引进随机变量 $t_i (i=1, 2, \dots, N)$:

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{Z_i} t_i,$$

于是根据 (5.2) ~ (5.4) 式, 有

$$\begin{aligned} E(\hat{Y}_{HH}) &= \frac{1}{n} \sum_{i=1}^N \frac{Y_i}{Z_i} E(t_i) = \sum_{i=1}^N Y_i = Y, \\ V(\hat{Y}_{HH}) &= \frac{1}{n^2} \left[\sum_{i=1}^N \frac{Y_i^2}{Z_i} V(t_i) + 2 \sum_{i=1}^N \sum_{j=1}^N \frac{Y_i Y_j}{Z_i Z_j} \text{Cov}(t_i, t_j) \right] \\ &= \frac{1}{n} \left[\sum_{i=1}^N \left(\frac{Y_i}{Z_i} \right)^2 Z_i (1 - Z_i) - 2 \sum_{i=1}^N \sum_{j=1}^N \frac{Y_i Y_j}{Z_i Z_j} Z_i Z_j \right] \\ &= \frac{1}{n} \left[\sum_{i=1}^N \frac{Y_i^2}{Z_i} (1 - Z_i) - 2 \sum_{i=1}^N \sum_{j=1}^N Y_i Y_j \right] \\ &= \frac{1}{n} \left[\sum_{i=1}^N \frac{Y_i^2}{Z_i} - \sum_{i=1}^N (Y_i^2 + 2 \sum_{j=1}^N Y_i Y_j) \right] \\ &= \frac{1}{n} \left[\sum_{i=1}^N \frac{Y_i^2}{Z_i} - Y^2 \right] \\ &= \frac{1}{n} \sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2. \end{aligned} \quad (5.11)$$

为证明 (5.9) 式, 注意到

$$\begin{aligned}
 1 - Z_i &= \sum_{j=1}^N Z_j - Z_i = \sum_{j \neq i}^N Z_j, \\
 \sum_{i=1}^N \frac{Y_i^2}{Z_i} (1 - Z_i) &= \sum_{i=1}^N \sum_{j \neq i}^N \frac{Y_i^2 Z_j}{Z_i} \\
 &= \sum_{i=1}^N \sum_{i < j}^N \left(\frac{Y_i^2 Z_j}{Z_i} + \frac{Y_j^2 Z_i}{Z_j} \right).
 \end{aligned}$$

将上式代入(5.11)式, 即有

$$\begin{aligned}
 V(\hat{P}_{HH}) &= \frac{1}{n} \sum_{i=1}^N \sum_{i < j}^N \left(\frac{Y_i^2 Z_j}{Z_i} + \frac{Y_j^2 Z_i}{Z_j} - 2Y_i Y_j \right) \\
 &= \frac{1}{n} \sum_{i=1}^N \sum_{i < j}^N Z_i Z_j \left(\frac{Y_i}{Z_i} - \frac{Y_j}{Z_j} \right)^2.
 \end{aligned}$$

为证明 $v(\hat{P}_{HH})$ 是 $V(\hat{P}_{HH})$ 的无偏估计, 注意到

$$\sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{P}_{HH} \right)^2 = \sum_{i=1}^n \left(\frac{y_i}{z_i} - Y \right)^2 - n(\hat{P}_{HH} - Y)^2,$$

于是有

$$\begin{aligned}
 n(n-1)E[v(\hat{P}_{HH})] &= E \left[\sum_{i=1}^n \left(\frac{y_i}{z_i} - Y \right)^2 \right] - nE(\hat{P}_{HH} - Y)^2 \\
 &= E \left[\sum_{i=1}^N \left(\frac{Y_i}{Z_i} - Y \right)^2 t_i \right] - nV(\hat{P}_{HH}) \\
 &= n \sum_{i=1}^N \left(\frac{Y_i}{Z_i} - Y \right)^2 Z_i - nV(\bar{P}_{HH}) \\
 &= n^2 V(\hat{P}_{HH}) - nV(\hat{P}_{HH}) \\
 &= n(n-1)V(\hat{P}_{HH}).
 \end{aligned}$$

从而

$$E[v(\hat{P}_{HH})] = V(\hat{P}_{HH}).$$

证明 2 由于多项抽样是 n 次独立地从同一总体中进行的抽样. 我们将 \hat{P}_{HH} 看成是从 $\left\{ \frac{Y_1}{Z_1}, \frac{Y_2}{Z_2}, \dots, \frac{Y_N}{Z_N} \right\}$ 这个“总体”独立地抽取的样本量为 n 的一个样本的平均数, 抽到 $\frac{Y_i}{Z_i}$ 的概率为 Z_i , 于是 $E(\hat{P}_{HH})$ 等于该“总体”的均值, 后者为 $\sum_{i=1}^N Z_i \frac{Y_i}{Z_i} = Y$, 从而 \hat{P}_{HH} 是无偏的. 又样本均值 \hat{P}_{HH} 的方差 $V(\hat{P}_{HH})$ 应为“总体”方差的 $\frac{1}{n}$, 而“总体”方差按定义恰好为 $\sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2$, 从而(5.7)式成立. “总体”方差可以用样本方差 $\frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{P}_{HH} \right)^2$ 估计, 后者是前者的无偏估计, 于是 $v(\hat{P}_{HH}) =$

$\frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{P}_{HH} \right)^2$ 是 $V(\hat{P}_{HH})$ 的无偏估计. ■

从定理 5.1 可以看出, 多项抽样的估计量及其方差估计都是十分简单的, 因此这种抽样在实际工作中应用相当广泛.

5.2.3 数值例子——职工人数的调查

例 5.2 表 5.2 是某系统全部 36 个单位的上一年职工人数 X_i 及当年职工人数 Y_i 的数据. 以 X_i 作为单位大小 M_i 的度量, 对单位进行 PPS 抽样, 估计全系统当年职工总人数 Y .

表 5.2 某系统各单位的上一年与当年职工人数

单位号	X_i	Y_i	单位号	X_i	Y_i
1	598	633	19	231	255
2	21	18	20	15	24
3	630	656	21	172	181
4	3012	3273	22	234	243
5	372	386	23	312	338
6	142	164	24	351	371
7	1027	1145	25	252	281
8	432	501	26	194	210
9	216	235	27	149	166
10	1698	1778	28	173	189
11	1570	1541	29	318	344
12	502	486	30	204	227
13	497	516	31	52	63
14	723	736	32	183	174
15	712	740	33	97	123
16	335	352	34	213	242
17	267	299	35	47	51
18	1658	1714	36	838	879

设实际共抽得 4 个样本, 每个样本的样本量均为 6, 4 个样本抽得的单位号码如下:

样本 I (4), (10), (23), (11), (13), (3);

样本 II (1), (14), (4), (16), (28), (9);

样本 III (12), (10), (36), (4), (24), (4);

样本 IV (14), (4), (18), (28), (11), (34).

对每个样本 α ($\alpha = 1, 2, 3, 4$), 用 Hansen-Hurwitz 估计量估计全

系统当年职工总数 Y , 为方便起见, 令

$$\bar{y}_\alpha = \frac{1}{n} \sum_{i=1}^n \frac{y_{\alpha i}}{x_{\alpha i}},$$

$$v(\bar{y}_\alpha) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_{\alpha i}}{x_{\alpha i}} - \bar{y}_\alpha \right)^2,$$

则

$$\hat{P}_\alpha = X \bar{y}_\alpha,$$

$$v(\hat{P}_\alpha) = X^2 v(\bar{y}_\alpha).$$

其中 $X = 18457$ 是上一年职工总数.

注意到将所有 4 个样本合在一起, 即得到一个 $n = 24$ 的新样本, 对这个样本同样可计算 \hat{P} 与 $v(\hat{P})$, 计算结果列成表 5.3.

表 5.3 对表 5.2 的总体进行 PPS 抽样, 对 Y 的估计及其精度

样本号 α	1	2	3	4	综合
\bar{y}_α	1.0463548	1.0772523	1.0490758	1.0652787	1.0594914
\hat{P}_α	19312.6	19882.8	19362.8	19661.8	19555.0
$\sqrt{v(\bar{y}_\alpha)}$	0.015584	0.007271	0.017742	0.019740	0.007823
$\sqrt{v(\hat{P}_\alpha)}$	287.6426	134.2003	327.4601	364.3344	144.3955

由于方差估计的不稳定, $\sqrt{v(\hat{P}_\alpha)}$ 的值在所抽得的 4 个样本中有较大的变化. 其中第二个样本的 $v(\hat{P}_\alpha)$ 明显低估了 $V(\hat{P}_\alpha)$. 随着样本量的增加, $v(\hat{P})$ 的稳定性也将提高. 对于综合样本, $\sqrt{v(\hat{P}_\alpha)}$ 的值应是比较可靠的. 另外, 根据表 5.2 的数据, 可计算实际的 $Y = 19583$, 当然也是综合样本的实际估计误差最小.

还有一种从所得 4 个样本中获得的综合估计及其方差估计的方法. 令

$$\hat{P}^* = \frac{1}{4} \sum_{\alpha=1}^4 \hat{P}_\alpha = 19555.0,$$

则
$$v(\hat{P}^*) = \frac{1}{4 \times 3} \sum_{\alpha=1}^4 (\hat{P}_\alpha - \hat{P}^*)^2 = 17879.81$$

也可作为 $v(\hat{P}^*)$ 的估计. 相应的标准差为

$$\sqrt{v(\hat{P}^*)} = 133.7154.$$

注意在计算 $v(\hat{P}^*)$ 时并没有用到 $v(\hat{P}_\alpha)$ 的数据, 是完全从样本估计量出发的. 这提供了一种复杂样本方差估计的方法. 关于这种方法, 详见第 9 章的讨论.

§ 5.3 不放回不等概率抽样与 Horvitz-Thompson 估计量

5.3.1 不放回不等概率抽样与包含概率

上节讨论的放回抽样, 虽然实施方便, 且总体参数估计及其方差估计也简单, 但有两个主要缺点:

1. 直观上没有必要将同一单元重复进行调查(观测), 因此, 放回抽样所得的样本的代表性比相应的不放回抽样差, 不易被实际调查者所接受.

2. 对同样的样本量, 放回抽样的精度比不放回抽样的差, 也即效率较低. 尽管不放回抽样在许多情况下方差不易求得, 但从简单随机抽样情形可知(参见 § 2.5), 不放回抽样的方差是相应的放回抽样方差的 $(N-n)/(N-1)$ 倍, 也即约为 $1-f$ 倍. 当 f 不能忽略时, 这个因素是需要认真考虑的.

在不放回不等概率抽样中, 总体中每个单元被包含到样本的概率即入样概率 $\pi_i = P_r(i)$ 及任意两个单元都包含到样本的概率 $\pi_{ij} = P_r(i, j)$ 起着十分重要的作用, 它们通称为包含概率(inclusion probabilities).

引理 5.1 对固定的 n , 包含概率满足:

$$1) \sum_{i=1}^N \pi_i = n; \quad (5.12)$$

$$2) \sum_{j \neq i}^N \pi_{ij} = (n-1)\pi_i; \quad (5.13)$$

$$3) \sum_{i=1}^N \sum_{j \neq i}^N \pi_{ij} = \frac{1}{2} n(n-1). \quad (5.14)$$

证明 1) 是显然的.

$$2) \sum_{j \neq i}^N \pi_{ij} = \sum_{j \neq i}^N P_r(i)P_r(j|i) = \pi_i \sum_{j \neq i}^N P_r(j|i) = (n-1)\pi_i.$$

$$3) \sum_{i=1}^N \sum_{j \neq i}^N \pi_{ij} = \frac{1}{2} \sum_{i=1}^N \sum_{j \neq i}^N \pi_{ij} = \frac{1}{2} (n-1) \sum_{i=1}^N \pi_i = \frac{1}{2} n(n-1). \quad \blacksquare$$

对不放回抽样, 我们最感兴趣的是每个单元入样概率与其大小 M_i 严格成比例的情形, 当 n 固定时, 记 $Z_i = M_i/M_0$. ($M_0 = \sum_{i=1}^N M_i$), 此时即有:

$$\pi_i = nZ_i. \quad (5.15)$$

以后我们将此种情形的抽样简称为严格的 π PS 抽样。

严格的 π PS 抽样, 不仅实施复杂, 而且由于此时 π_{ij} 不易求得, 方差估计也很复杂。特别是当 n 比较大时, 有时简直不可能。一个极端的情形是当 $n=N$, 此时所有单元都入样, 从而必然是等概率的。事实上, 严格的 π PS 抽样只有在 $n=2$ 时才有一些比较简单且实用的方法。对一般的 $n>2$, 严格的 π PS 抽样相当复杂。但有几种非严格的方法可供使用。Brewer 与 Hanif(1983) 总结了 50 种不放回的不等概率抽样, 但能够在实际中方便应用的却为数不多。

5.3.2 Horvitz-Thompson 估计量及其性质

对不放回不等概率抽样, π_i 是第 i 单元的包含概率, Horvitz 与 Thompson 在 1952 年提出了对总体总和 Y 的以下估计量:

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}, \quad (5.16)$$

此后我们称 \hat{Y}_{HT} 为 Horvitz-Thompson 估计量。

与放回抽样情形的 Hansen-Hurwitz 估计量相类似, 由于 $\pi_i(Z_i)$ 只是第 i 个单元的属性, 故每个观测值 y_i 在估计量中的权是不随该单元在何时用何种方式抽得而改变的。

定理 5.2 若 $\pi_i > 0 (i=1, 2, \dots, N)$, 则 Horvitz-Thompson 估计 \hat{Y}_{HT} 是 Y 的无偏估计, 其方差为:

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \frac{1-\pi_i}{\pi_i} Y_i^2 + 2 \sum_{i=1}^N \sum_{j=1}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j, \quad (5.17)$$

当 n 固定时, 又有

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{j=1}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2. \quad (5.18)$$

证明 引进随机变量

$$a_i = \begin{cases} 1, & \text{若第 } i \text{ 个单元入样} \\ 0, & \text{否则} \end{cases} \quad (i=1, 2, \dots, N), \quad (5.19)$$

■

$$E(a_i) = \pi_i, \quad (5.20)$$

$$V(a_i) = \pi_i(1-\pi_i), \quad (5.21)$$

$$\text{Cov}(a_i, a_j) = \pi_{ij} - \pi_i \pi_j \quad (i \neq j). \quad (5.22)$$

此时, \hat{Y}_{HT} 可表成:

$$\hat{Y}_{HT} = \sum_{i=1}^N a_i \frac{Y_i}{\pi_i}, \quad (5.23)$$

于是

$$\begin{aligned} E(\hat{Y}_{HT}) &= \sum_{i=1}^N \frac{Y_i}{\pi_i} E(a_i) = \sum_{i=1}^N Y_i = Y, \\ V(\hat{Y}_{HT}) &= \sum_{i=1}^N \frac{Y_i^2}{\pi_i^2} V(a_i) + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{Y_i Y_j}{\pi_i \pi_j} \text{Cov}(a_i, a_j) \\ &= \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} Y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j. \end{aligned}$$

当 n 固定时, 根据引理 5.1, 有

$$\begin{aligned} \sum_{j \neq i} (\pi_i \pi_j - \pi_{ij}) &= \pi_i \sum_{j \neq i} \pi_j - \sum_{j \neq i} \pi_{ij} \\ &= \pi_i (n - \pi_i) - (n - 1) \pi_i = \pi_i (1 - \pi_i). \end{aligned}$$

从而

$$\begin{aligned} \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} Y_i^2 &= \sum_{i=1}^N \sum_{j \neq i}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} \right)^2 \\ &= \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) \left[\left(\frac{Y_i}{\pi_i} \right)^2 + \left(\frac{Y_j}{\pi_j} \right)^2 \right], \end{aligned}$$

故

$$\begin{aligned} V(\hat{Y}_{HT}) &= \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) \left[\left(\frac{Y_i}{\pi_i} \right)^2 + \left(\frac{Y_j}{\pi_j} \right)^2 - 2 \frac{Y_i Y_j}{\pi_i \pi_j} \right] \\ &= \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2. \end{aligned}$$

为获得方差估计, 我们有以下的定理:

定理 5.3 若所有的 $\pi_i > 0$, $\pi_{ij} > 0$ ($i, j = 1, 2, \dots, N; i \neq j$), 则

$$v(\hat{Y}_{HT}) = \sum_{i=1}^n \frac{1 - \pi_i}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j \quad (5.24)$$

是 $V(\hat{Y}_{HT})$ 的无偏估计. 又当 n 固定时,

$$v_{YGS}(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (5.25)$$

也是 \hat{Y}_{HT} 的无偏估计.

证明 仍引进随机变量 a_i ($i = 1, 2, \dots, N$), 则

$$v(\hat{Y}_{HT}) = \sum_{i=1}^N a_i \frac{1 - \pi_i}{\pi_i^2} Y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N a_i a_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} Y_i Y_j,$$

从而

$$E[v(\hat{Y}_{HT})] = \sum_{i=1}^N \frac{1 - \pi_i}{\pi_i} Y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j = V(\hat{Y}_{HT}).$$

当 n 固定时, $v_{YGS}(\hat{Y}_{HT})$ 可表成:

$$v_{YGS}(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{j=1}^N a_i a_j \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2,$$

$$\text{从而 } E[v_{YGS}(\hat{Y}_{HT})] = \sum_{i=1}^N \sum_{j=1}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 = V(\hat{Y}_{HT}).$$

在实际问题中, $v(\hat{Y}_{HT})$ 与 $v_{YGS}(\hat{Y}_{HT})$ 都有可能取负值, 这样的方差估计当然还不够理想. 相对来说, $v_{YGS}(\hat{Y}_{HT})$ 比 $v(\hat{Y}_{HT})$ 稳定, 取负值的可能性小得多. Vijayan(1975)证明了在 $n=2$ 的情形, $v_{YGS}(\hat{Y}_{HT})$ 是唯一可能的恒为非负的方差估计. $v_{YGS}(\cdot)$ 是由 Sen(1953) 提出并由 Yates 与 Grundy(1953) 首次用于 n 固定的情形, 故它通常称为 Yates-Grundy-Sen 估计量.

§ 5.4 几种严格的不放回 π PS 抽样方法

本节介绍几种比较实用的严格不放回 π PS 抽样方法. 正如 § 5.1 所指出的, 这里的“严格不放回 π PS”是指样本量 n 固定, 严格不放回, 包含概率 π_i 严格与单元大小成比例, 即 $\pi_i = nZ_i$. 我们先介绍适用于 $n=2$ 的方法, 然后讨论适用于 $n>2$ 的一般方法.

5.4.1 $n=2$ 的情形

对于 $n=2$ 的情形, 在总体(或层)中仅需抽 2 个单元, 为了保证是不放回的, 故一般采用逐个抽取法. 先按给定的概率在总体中抽取第一个样本单元, 然后在剩下的单元中再按给定的概率抽取第二个样本单元. 上述概率要保证最终得到的样本单元被抽到的概率 $\pi_i = 2Z_i$, 此时包含概率 π_{ij} 即是包含单元 i 与 j 的样本被抽中的概率.

一、Brewer 方法(1963)

若对所有的 i , 都有 $Z_i < \frac{1}{2}$, 则两个样本单元的抽取方法是: 第一个单元按与 $\frac{Z_i(1-Z_i)}{1-2Z_i}$ 成比例的概率抽取; 第二个单元则在剩下的 $N-1$ 个单元中按与 Z_j 成正比的概率抽取. 下面证明按这种抽样方法, 有 $\pi_i = 2Z_i$.

令

$$D = \sum_{i=1}^N \frac{Z_i(1-Z_i)}{1-2Z_i} = \frac{1}{2} \sum_{i=1}^N \left(Z_i + \frac{Z_i}{1-2Z_i} \right) = \frac{1}{2} \left(1 + \sum_{i=1}^N \frac{Z_i}{1-2Z_i} \right), \quad (5.26)$$

于是第一个样本单元抽到单元 i 的概率为

$$\frac{Z_i(1-Z_i)}{D(1-2Z_i)}, \quad (5.27)$$

而第一个样本单元抽到 j , 第二个样本单元抽到 i 的概率应为

$$\frac{Z_i}{1-Z_j} \cdot \frac{Z_j(1-Z_j)}{D(1-2Z_j)} = \frac{Z_i Z_j}{D(1-2Z_j)}. \quad (5.28)$$

因此, 第二次抽到单元 i 的概率, 即是上述概率对 $j \neq i$ 的和. 于是在两次抽样中, 抽到单元 i 的概率 π_i 应为

$$\begin{aligned} \pi_i &= \frac{Z_i(1-Z_i)}{D(1-2Z_i)} + \sum_{j \neq i}^N \frac{Z_i Z_j}{D(1-2Z_j)} = \frac{Z_i}{D} \left[1 + \frac{Z_i}{1-2Z_i} + \sum_{j \neq i}^N \frac{Z_j}{1-2Z_j} \right] \\ &= \frac{Z_i}{D} \left[1 + \sum_{j=1}^N \frac{Z_j}{1-2Z_j} \right] - \frac{Z_i}{D} (2D) = 2Z_i. \end{aligned} \quad (5.29)$$

根据 (5.28) 式, 可以计算由单元 i 及 j 组成的样本被抽中的概率:

$$\begin{aligned} \pi_{ij} &= \frac{Z_i Z_j}{D} \left(\frac{1}{1-2Z_i} + \frac{1}{1-2Z_j} \right) = \frac{2Z_i Z_j}{D} \cdot \frac{1-Z_i-Z_j}{(1-2Z_i)(1-2Z_j)} \\ &= \frac{4Z_i Z_j (1-Z_i-Z_j)}{(1-2Z_i)(1-2Z_j) \left[1 + \sum_{i=1}^N \frac{Z_i}{1-2Z_i} \right]}. \end{aligned} \quad (5.30)$$

于是根据 Horvitz-Thompson 估计, 总体总和 Y 的估计为

$$\hat{Y}_B = \frac{y_i}{\pi_i} + \frac{y_j}{\pi_j} = \frac{1}{2} \left(\frac{y_i}{z_i} + \frac{y_j}{z_j} \right). \quad (5.31)$$

根据 (5.30) 及 (5.25) 式, 即可得到 $V(\hat{Y}_B)$ 的 Yates Grundy Sen 估计 $v_{YGS}(\hat{Y}_B)$, 注意到

$$\begin{aligned} &(1-2Z_i)(1-2Z_j) \left[1 + \frac{Z_i}{1-2Z_i} + \frac{Z_j}{1-2Z_j} \right] \\ &= (1-2Z_i)(1-2Z_j) + Z_i(1-2Z_j) + Z_j(1-2Z_i) \\ &= 1 - Z_i - Z_j, \end{aligned}$$

从而

$$\frac{1 - Z_i - Z_j}{(1-2Z_i)(1-2Z_j) \left(1 + \sum_{i=1}^N \frac{Z_i}{1-2Z_i} \right)} < 1,$$

$$\pi_{ij} < 4Z_i Z_j = \pi_i \pi_j. \quad (5.32)$$

从而 $v_{YGS}(\hat{Y}_B)$ 恒为正.

二、Durbin 方法 (1967)

第一个样本单元以概率 Z_i 抽取, 设第 i 个单元入样; 第二个样本单元以与 $Z_j \left(\frac{1}{1-2Z_i} + \frac{1}{1-2Z_j} \right)$ 成正比的概率抽取. 为计算 π_i 与 π_{ij} , 令

$$\begin{aligned}
 D' &= \sum_{j=1}^N Z_j \left(\frac{1}{1-2Z_j} + \frac{1}{1-2Z_j} \right) \\
 &= \frac{1}{1-2Z_1} + \sum_{j=2}^N \frac{Z_j}{1-2Z_j} \\
 &= 1 + \sum_{j=1}^N \frac{Z_j}{1-2Z_j} = 2D.
 \end{aligned} \tag{5.33}$$

于是

$$\begin{aligned}
 \pi_i &= Z_i + \sum_{j \neq i}^N Z_j Z_i \left(\frac{1}{1-2Z_j} + \frac{1}{1-2Z_j} \right) \\
 &= Z_i + Z_i \frac{D'}{D} = 2Z_i,
 \end{aligned} \tag{5.34}$$

$$\begin{aligned}
 \pi_{ij} &= 2Z_i Z_j \left(\frac{1}{1-2Z_i} + \frac{1}{1-2Z_j} \right) / 2D \\
 &= \frac{Z_i Z_j}{D} \left(\frac{1}{1-2Z_i} + \frac{1}{1-2Z_j} \right).
 \end{aligned} \tag{5.35}$$

这与 Brewer 方法的 π_{ij} 相等, 因此 Durbin 方法实际上与 Brewer 方法是等价的.

三、Hanurav 方法(1937)

两个样本单元的抽取按以下步骤进行:

1) 按 Z_i 的递增顺序将总体单元重新排列:

$$Z_1 \leq Z_2 \leq \cdots \leq Z_N.$$

2) 以

$$\delta = \frac{2(1-Z_N)(Z_N - Z_{N-1})}{1 - Z_N - Z_{N-1}} \tag{5.36}$$

为成功概率作 Bernoulli 试验.

3) 若 2) 中的试验成功, 则第 N 个单元入样, 再以与 Z_i 成比例的概率抽取另一个单元.

4) 若 2) 中的试验失败, 则令

$$Z'_N = Z_{N-1}, \quad Z'_k = Z_k \quad (k=1, 2, \dots, N-1). \tag{5.37}$$

$$Z'_i = \frac{Z'_i}{\sum_{k=1}^N Z'_k} \quad (i=1, 2, \dots, N), \tag{5.38}$$

以下面的工作概率抽取第一个单元:

$$\alpha_1 = 2Z'_1,$$

$$\alpha_i = 2Z'_i - \frac{\alpha_1}{N-1} - \frac{\alpha_2}{N-2} - \cdots - \frac{\alpha_{i-1}}{N-i+1} \quad (i=2, \dots, N). \tag{5.39}$$

在抽得的单元顺序后面(注意: 此时 α_N 必为 0)的单元中, 以等概率抽取一个单元作为第二个样本单元.

按此种抽样方法, 可证明

$$\begin{aligned} \sigma_i &= 2Z_i, \\ \sigma_{ij} &= \begin{cases} \frac{\alpha_i}{N-1}(1-\delta) & (i < j \neq N); \\ \frac{Z_i \delta}{1-Z_N} + \frac{\alpha_i(1-\delta)}{N-i} & (i < j = N). \end{cases} \end{aligned} \quad (5.40)$$

四、Narain 方法(1951)

计算一组工作概率 Z_i^* , 用此概率抽取第一个单元, 然后在剩下的 $N-1$ 个单元中以与 Z_i^* 成比例的概率抽取第二个单元, 此时

$$\sigma_i = Z_i^* + \sum_{j \neq i}^N \frac{Z_j^*}{1-Z_j^*} \cdot Z_j^*, \quad (5.41)$$

$$\sigma_{ij} = Z_i^* Z_j^* \left[\frac{1}{1-Z_i^*} + \frac{1}{1-Z_j^*} \right]. \quad (5.42)$$

为保证方法是严格 π PS 的, Z_i^* 必须满足使每个 $\sigma_i = 2Z_i$, 因此 Z_i^* 需用迭代法进行计算, 读者可参阅 Brewer 与 Hanif 的书《Sampling with Unequal Probabilities》(1983)中的附录 A.

五、Fellegi 方法(1963)

这个方法与 Narain 方法类似, 直接以 Z_i 的概率抽取第一个样本单元, 不放回, 再以与 Z_i 成比例的概率 Z_i^* 抽取第二个样本单元, Z_i^* 也需用迭代法求出.

5.4.2 $n > 2$ 的情形

在实际应用中, $n=2$ 的情形通常用在先对单元分层, 在每层内抽取 2 个单元的情况. 对于一般的 n , 必须用本段介绍的方法, 不过此时包含概率 σ_{ij} 的计算通常极为复杂.

一、Rao-Sampford 重抽法(1965, 1967)

这种方法是先以 Z_i 的概率抽取第一个样本单元, 然后以与

$$\lambda_i = \frac{Z_i}{1-nZ_i} \quad (5.43)$$

成比例的概率依次放回抽取 $n-1$ 个单元(设所有的 $Z_i < \frac{1}{n}$). 在此过程中, 一旦有单元重复被抽中, 则全部放弃已抽到的单元, 再重抽, 直到抽中的 n 个单元都不同为止. 此时可证明

$$\pi_i = nZ_i, \quad (5.44)$$

$$\pi_{ij} = \frac{2K_n \lambda_i \lambda_j}{n(n-1)} \sum_{t=2}^n \left[\frac{t - n(Z_i + Z_j)}{n^{t-2}} L_{n-t}(\tilde{i}, \tilde{j}) \right]. \quad (5.45)$$

其中

$$K_n = \left[\sum_{t=1}^n \frac{t L_{n-t}}{n^t} \right]^{-1},$$

而

$$L_0 = 1, \quad L_m = \sum_{s \in (m)} \lambda_{i_s} \lambda_{i_{s_1}} \cdots \lambda_{i_{s_m}} \quad (m=1, \dots, n).$$

这里的 $\sum_{s \in (m)}$ 是对总体中所有可能的互不相同的 m 个单元组成的样本求和, $L_m(\tilde{i}, \tilde{j})$ 是对除去 i, j 两个单元的子总体中相应的 L_m 值.

特别地, 当 $n=2$ 时,

$$\begin{aligned} \pi_{ij} &= \frac{2K_2 \lambda_i \lambda_j}{2} [2(1 - Z_i - Z_j)] \\ &= 2 \left[\frac{1}{2} \left(1 + \sum_{i=1}^N \frac{Z_i}{1 - 2Z_i} \right)^{-1} \left(\frac{Z_i}{1 - 2Z_i} \right) \left(\frac{Z_j}{1 - 2Z_j} \right) (1 - Z_i - Z_j) \right] \\ &= \frac{4Z_i Z_j (1 - Z_i - Z_j)}{(1 - 2Z_i)(1 - 2Z_j) \left[1 + \sum_{i=1}^N \frac{Z_i}{1 - 2Z_i} \right]}. \end{aligned}$$

这与 Brewer 方法和 Durbin 方法是一致的.

Rao Samford 方法的优点是可以通过计算机得到精确的 π_{ij} 的值. 但由于计算量大, 一般也只适用于 n 不很大的情形. 为此有人给出了这种方法的求 π_{ij} 的近似公式, 可以较大程度地减少计算量.

二、水野 (Midzuno) 方法 (见 Horvitz 与 Thompson (1952) 的报告)

水野方法是一种逐个抽取的方法, 应用起来较为方便. 它的步骤是以概率

$$Z_i^* = \frac{n(N-1)Z_i}{N-n} = \frac{n-1}{N-n} \quad (i=1, 2, \dots, N) \quad (5.46)$$

抽取第一个样本单元, 在剩下的 $N-1$ 个单元中不放回地等概率抽取 $n-1$ 个单元. 容易验证所有 Z_i^* 之和等于 1. 但为了保证对所有的 i 都有 $Z_i^* \geq 0$, 则需要每个单元的大小

$$M_i \geq \frac{(n-1)M_0}{n(N-1)}, \quad (5.47)$$

为做到这一点, 必须避免 M_i 相差过大. 这可以通过分层来达到, 即将大小相仿的单元分在同一层. 在 (5.47) 式成立的条件下, 有

$$\pi_i = nZ_i,$$

$$\sigma_{ij} = \frac{n-1}{N-1} \left[\frac{N-n}{N-2} (Z_i^* + Z_j^*) + \frac{n-2}{N-n} \right], \quad (5.48)$$

三、Brewer 方法(1963)

这也是一种逐个不放回抽取法, 是 $n=2$ 时的 Brewer 方法的推广. 此方法是以与 $\frac{Z_i(1-Z_i)}{1-nZ_i}$ 成比例的概率抽取第一个单元. 在第 r 次抽取时, 以与

$$\frac{Z_i(1-Z_i)}{1-(n-r+1)Z_i} \quad (5.49)$$

成比例的概率从尚未入样的单元中抽取一个单元. 这也是一种严格的 π PS 抽样, 但 σ_{ij} 的公式相当复杂, 不过有递推公式可以使用.

§ 5.5 其他不放回抽样方法及其相应的估计量

从上节可以看到, 当 $n>2$ 时, 严格的 π PS 抽样方法不论是方法本身还是方差估计(表现在 σ_{ij} 的计算上)都是很复杂的. 在实用中有时采用一些非严格的抽样方法. 这里的“非严格”是指以下任何一种情况: π_i 不严格等于 nZ_i ; 不是严格不放回的; 样本量 n 不固定从而是随机的. 对于非严格的 π PS 抽样, 有时需要采用特殊的估计量. 本节讨论几种需要用特殊估计量的非严格 π PS 抽样方法.

5.5.1 Yates-Grundy 逐个抽取法及 Das-Raj-Murthy 估计量

Yates-Grundy 逐个抽取法(1953)是逐个不放回地抽取单元, 每次抽取皆按当时未入样的单元的 Z_i 成比例的概率抽取. 即第一个样本单元按 Z_i 的概率抽取, 设第 i 个单元入样; 第二个样本单元按 $Z_j/(1-Z_i)$ 的概率在其余 $N-1$ 个单元中抽取, 设第 j 个单元入样; 第三个样本单元则按 $Z_k/(1-Z_i-Z_j)$ 的概率在剩下的 $N-2$ 个单元中抽取; 以此类推, 直至抽够 n 个单元为止. 按这种方法, π_i 显然不是严格地与 Z_i 成比例. 但由于在不放回不等概率抽样中, 这种抽样是最自然的, 也是最简单的方法, 故在实际中得到相当广泛的使用.

对于上述抽样, 由于 π_i 不易计算, 故不能用 Horvitz-Thompson 估计. 对此, Das(1951)最早提出以下估计方法. 设 y_1, y_2, \dots, y_n 是按抽中顺序排列的样本单元(的指标值), 相应的 Z 值为 z_1, z_2, \dots, z_n , 令

$$\begin{cases} t'_1 = \frac{y_1}{z_1}, \\ t'_2 = \frac{1 - z_1}{z_1 z_2} \cdot \frac{y_2}{N-1}, \\ \dots\dots\dots \\ t'_n = \frac{\prod_{i=1}^{n-1} \left(1 - \sum_{j=1}^i z_j\right) y_n}{\prod_{i=1}^n z_i \prod_{i=1}^{n-1} (N-i)}. \end{cases} \quad (5.50)$$

事实上, 每个 t'_i 都是总体总和 Y 的一个无偏估计, 我们取它们的平均数

$$\hat{Y}_D = \frac{1}{n} \sum_{i=1}^n t'_i, \quad (5.51)$$

则它也是 Y 的无偏估计. 但因为 t'_i 彼此是相关的, 故 \hat{Y}_D 的方差计算很困难. 为此, Raj(1956)修正了 Das 的估计量, 令

$$\begin{cases} t_1 = \frac{y_1}{z_1}, \\ t_2 = y_1 + \frac{y_2}{z_2} (1 - z_1), \\ \dots\dots\dots \\ t_n = \sum_{i=1}^{n-1} y_i + \frac{y_n}{z_n} \left(1 - \sum_{i=1}^{n-1} z_i\right). \end{cases} \quad (5.52)$$

每个 t_i 仍是 Y 的无偏估计, 但彼此不相关, 因此, Raj 估计量

$$\hat{Y}_R = \frac{1}{n} \sum_{i=1}^n t_i \quad (5.53)$$

不仅是无偏的, 且能求得它的方差表达式:

$$V(\hat{Y}_R) = \frac{1}{2n^2} \sum_{i=1}^N \sum_{j=1}^N Z_i Z_j \left[1 + \sum_{r=2}^n Q_{ij}(r-1) \right] \left(\frac{Y_i}{Z_i} - \frac{Y_j}{Z_j} \right)^2, \quad (5.54)$$

式中 $Q_{ij}(r-1)$ 表示总体中第 i, j 个单元在前 $r-1$ 次抽取时有一个或都没有被抽到的概率. (5.54) 式尽管形式复杂, 但鉴于 t_i 的不相关性, 它有以下简单的无偏估计量:

$$v(\hat{Y}_R) = \frac{1}{n(n-1)} \sum_{i=1}^n (t_i - \hat{Y}_R)^2, \quad (5.55)$$

也即作为 t_i 的样本平均数 \hat{Y}_R 的方差, 可用 t_i 的样本方差的 $1/n$ 进行估计.

Raj 估计量与 Das 估计量一样, 都与单元的入样顺序有关, 这当

然很不方便, Murthy(1957)证明了上述估计量可以通过考虑已知样本中的单元的所有可能的置换而得到改进. 他提出以下与入样次序无关的估计量:

$$\hat{P}_M = \frac{\sum_{i=1}^n P_r(S_i|i)y_i}{P_r(S)}. \quad (5.56)$$

式中 $P_r(S_i|i)$ 表示在首次抽取时抽中第 i 个单元的条件下抽到最终样本 S 的条件概率, 而 $P_r(S)$ 则是抽中样本 S 的无条件概率. 下面证明 Murthy 估计量 \hat{P}_M 是无偏的. 为此先证明对任意的 i , 有

$$\sum_{S_i} P_r(S_i|i) = 1. \quad (5.57)$$

式中求和是对所有第一次抽样抽到单元 i 的样本求的. 对 $n=2$, 设单元 j 为另一样本单元, 则

$$P_r(S_i|i) = \frac{Z_i}{1-Z_i},$$

故

$$\sum_{S_i} P_r(S_i|i) = \sum_{j \neq i}^N \frac{Z_j}{1-Z_i} = 1.$$

对 $n=3$, 设单元 j, k 分别为第二、三次被抽中的样本单元, 则

$$\sum_{S_i} P_r(S_i|i) = \sum_{j \neq i}^N \sum_{k \neq i, j}^N \frac{Z_i Z_k}{(1-Z_i)(1-Z_i-Z_j)} = \sum_{j \neq i}^N \frac{Z_j}{1-Z_i} = 1.$$

同样可证明, 对一般的 n , (5.57) 式成立. 于是

$$\begin{aligned} E(\hat{P}_M) &= \sum_S P_r(S) \hat{P}_M = \sum_S \sum_{i=1}^n P_r(S_i|i)y_i \\ &= \sum_{i=1}^n \sum_{S_i} P_r(S_i|i)Y_i = \sum_{i=1}^n Y_i = Y. \end{aligned} \quad (5.58)$$

当 $n=2$ 时, 记 $S=(i, j)$, 即 i, j 是两个样本单元, 则

$$\begin{aligned} P_r(S_i|i) &= \frac{Z_i}{1-Z_i}, \quad P_r(S_j|j) = \frac{Z_j}{1-Z_j}, \\ P_r(S) &= \pi_{ij} = Z_i P_r(S_i|i) + Z_j P_r(S_j|j) \\ &= \frac{Z_i Z_j (2 - Z_i - Z_j)}{(1-Z_i)(1-Z_j)}. \end{aligned} \quad (5.59)$$

从而

$$\hat{P}_M = \frac{1}{2-Z_i-Z_j} \left[(1-Z_i) \frac{Y_i}{Z_i} + (1-Z_j) \frac{Y_j}{Z_j} \right], \quad (5.60)$$

$$\begin{aligned}
V(\hat{P}_M) &= E(\hat{P}_M^2) - [E(\hat{P}_M)]^2 \\
&= \sum_S P_r(S) \hat{P}_M^2 - Y^2 \\
&= \sum_{i=1}^N \sum_{j=1}^N \frac{Z_i Z_j (2 - Z_i - Z_j)}{(1 - Z_i)(1 - Z_j)} \hat{P}_M^2 - Y^2 \\
&= \sum_{i=1}^N \sum_{j=1}^N \frac{Z_i Z_j (1 - Z_i - Z_j)}{2 - Z_i - Z_j} \left(\frac{Y_i}{Z_i} - \frac{Y_j}{Z_j} \right)^2, \quad (5.61)
\end{aligned}$$

与多项抽样的 Hansen-Hurwitz 估计量的方差公式(5.9 式)比较, 由于

$$\frac{1 - Z_i - Z_j}{2 - Z_i - Z_j} < \frac{1}{2},$$

故 $V(\hat{P}_M)$ 小于 $V(\hat{P}_{HH})$.

可以直接验证, 在 $n=2$ 的情形, $V(\hat{P}_M)$ 的一个无偏估计为

$$v(\hat{P}_M) = \frac{(1 - Z_i)(1 - Z_j)(1 - Z_i - Z_j)}{(2 - Z_i - Z_j)^2} \left(\frac{Y_i}{Z_i} - \frac{Y_j}{Z_j} \right)^2. \quad (5.62)$$

只要所有的 $Z_i < \frac{1}{2}$, 则 $v(\hat{P}_M)$ 恒为正.

如果用惯用的记号重新将两个样本单元的编号记为 1 与 2, 则 (5.60) 与 (5.62) 式可写成:

$$\hat{P}_M = \frac{1}{2 - z_1 - z_2} \left[(1 - z_2) \frac{y_1}{z_1} + (1 - z_1) \frac{y_2}{z_2} \right], \quad (5.63)$$

$$v(\hat{P}_M) = \frac{(1 - z_1)(1 - z_2)(1 - z_1 - z_2)}{(2 - z_1 - z_2)^2} \left(\frac{y_1}{z_1} - \frac{y_2}{z_2} \right)^2. \quad (5.64)$$

对于 $n > 2$ 的一般情形, \hat{P}_M 的方差具有以下形式:

$$V(\hat{P}_M) = \sum_{i=1}^N \sum_{j=1}^N \left[1 - \sum_S^* \frac{P_r(S|i)P_r(S|j)}{P_r(S)} \right] Z_i Z_j \left(\frac{Y_i}{Z_i} - \frac{Y_j}{Z_j} \right)^2, \quad (5.65)$$

式中 \sum_S^* 是对所有包含单元 i, j 的样本求和, $P_r(S|i)$ 、 $P_r(S|j)$ 分别是第一次抽中的单元 i 或单元 j , 最终抽中样本 S 的条件概率. $V(\hat{P}_M)$ 的一个无偏估计是

$$\begin{aligned}
v(\hat{P}_M) &= \frac{1}{[P_r(S)]^2} \left\{ \sum_{i=1}^n \sum_{j=1}^n [P_r(S)P_r(S|i, j) - P_r(S|i)P_r(S|j)] \right. \\
&\quad \left. \times z_i z_j \left(\frac{y_i}{z_i} - \frac{y_j}{z_j} \right)^2 \right\}, \quad (5.66)
\end{aligned}$$

其中 $P_r(S|i, j)$ 是在前两次抽中单元 i 与 j (不计顺序) 的条件下抽到最终样本 S 的条件概率. $v(\hat{P}_M)$ 的计算必须借助计算机编制专门程序才能进行, 而且计算量随着 n 的增大也急剧增大.

5.5.2 Rao-Hartley-Cochran 方法及其估计量

这是由 Rao Hartley-Cochran 于 1962 年提出的一种简单而实用的方法. 将总体中的单元随机地分成 n 组, 每组的单元数记为 N_1, N_2, \dots, N_n , 在每组中按与 Z_i 成比例的概率抽取一个单元入样, 即若 Z_g^* 是第 g 组 N_g 个单元 Z_i 值的总和, 则按 Z_i/Z_g^* 概率抽取. 将被抽到的单元的观测值记为 y_g , 相应的 Z 值记为 z_g , 则 Rao Hartley-Cochran 估计量定义为:

$$\hat{P}_{RHC} = \sum_{g=1}^n Z_g^* \frac{y_g}{z_g}. \quad (5.67)$$

由于 Z_g^* 并不相等, 因此就整体而言, 这种抽样并不是严格 π PS 的, 但是在每一组中, 抽样是严格 PPS 或 π PS 的 (由于 $n_g = 1$, 故无所谓放回或不放回), 从而对组总和 Y_g 的估计

$$\hat{P}_g = \frac{y_g}{z_g/Z_g^*} \quad (5.68)$$

是无偏的, 因而 \hat{P}_{RHC} 是总体总和 Y 的无偏估计.

至于 \hat{P}_{RHC} 的方差, 它有两个来源: 一是由于分组的随机性, 二是由于组内的抽样. 根据引理 3.1 的一般结果, 有

$$V(\hat{P}_{RHC}) = E_1[V_2(\hat{P}_{RHC})] + V_1[E_2(\hat{P}_{RHC})]. \quad (5.69)$$

其中 E_1, V_1 分别表示随机分组的期望与方差, E_2, V_2 分别表示在固定分组条件下组内抽样的期望与方差. 前面已论述了 \hat{P}_g 与 \hat{P}_{RHC} 的无偏性, 因此 (5.69) 中的第二项等于 0, 而根据 (5.9) 式, 对每一 \hat{P}_g , 有 (注意此时 $n_g = 1$)

$$\begin{aligned} V_2(\hat{P}_g) &= V(\hat{P}_g) = \sum_{i=1}^{N_g} Z_i Z_i \left(\frac{Y_i}{Z_i} - \frac{Y}{Z_g} \right)^2, \\ E_1[V_2(\hat{P}_g)] &= \frac{N_g(N_g-1)}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N Z_i Z_j \left(\frac{Y_i}{Z_i} - \frac{Y_j}{Z_j} \right)^2 \\ &= \frac{N_g(N_g-1)}{N(N-1)} \left(\sum_{i=1}^N \frac{Y_i^2}{Z_i} - Y^2 \right). \end{aligned}$$

而
$$\hat{P}_{RHC} = \sum_{g=1}^n \hat{P}_g,$$

因此
$$V(\hat{P}_{RHC}) = E_1[V_2(\hat{P}_{RHC})] = \sum_{g=1}^n E_1[V_2(\hat{P}_g)]$$

$$= \frac{\left(\sum_{g=1}^n N_g^2 - N \right)}{N(N-1)} \left(\sum_{i=1}^N \frac{Y_i^2}{Z_i} - Y^2 \right). \quad (5.70)$$

与从总体中按放回 PPS 抽样的 Hansen-Hurwitz 估计量的方差 $V(\hat{P}_{HH})$ 比较, (5.70) 式可写成

$$V(\hat{P}_{RHO}) = \frac{n \left(\sum_{g=1}^n N_g^2 - N \right)}{N(N-1)} V(\hat{P}_{HH}). \quad (5.71)$$

上式表明 $V(\hat{P}_{RHO})$ 可表成 $V(\hat{P}_{HH})$ 乘上一个因子的形式. 若 $\frac{N}{n} = R$ 是一整数, 则取 $N_g = R$ 即能使 $V(\hat{P}_{RHO})$ 达到极小:

$$V_{\min}(\hat{P}_{RHO}) = \left(1 - \frac{n-1}{N-1} \right) V(\hat{P}_{HH}) = \frac{N}{N-1} \frac{n}{n-1} V(\hat{P}_{HH}). \quad (5.72)$$

若 $N = nR + k$ (R 为整数, $k < n$), 则使 (5.70) 或 (5.71) 达到极小的分组是取 k 组的大小为 $R+1$, 其余 $n-k$ 组的大小为 R , 此时

$$V_{\min}(\hat{P}_{RHO}) = \left[1 - \frac{n-1}{N-1} + \frac{k(n-k)}{N(N-1)} \right] V(\hat{P}_{HH}). \quad (5.72')$$

至于方差的估计, 可以证明 (5.70) 的一个无偏估计是

$$v(\hat{P}_{RHO}) = \frac{\sum_{g=1}^n N_g^2 - N}{N^2 - \sum_{g=1}^n N_g^2} \sum_{g=1}^n Z_g^* \left(\frac{y_g}{z_g} - \hat{P}_{RHO} \right)^2. \quad (5.73)$$

而 (5.72) 式的一个无偏估计是

$$v_{\min}(\hat{P}_{RHO}) = \frac{N^2 + k(n-k)}{N^2(n-1)} \frac{Nn}{k(n-k)} \sum_{g=1}^n Z_g^* \left(\frac{y_g}{z_g} - \hat{P}_{RHO} \right)^2. \quad (5.74)$$

5.5.3 Poisson 抽样

Hajek (1964) 设计了一种严格 π PS、严格不放回, 但 n 不事先固定的抽样方法, 称为 Poisson 抽样. 对每个总体单元赋予一个入样概率 π_i , 使 $\pi_i/Z_i = \nu$, 其中 ν 是一常数. 以 π_i 为成功概率, 作一次 Bernoulli 试验, 若试验成功, 则相应的单元入样. 共作 N 次这样的试验, 实际入样的单元数即样本量 n 是一个随机变量. 显然, 有

$$E(n) = \sum_{i=1}^N \pi_i = \nu. \quad (5.75)$$

总体总和 Y 的估计有两种方法: 一种仍是采用 Horvitz Thompson 估计, 即

$$\hat{Y}_{PS} = \sum_{i=1}^n \frac{y_i}{\pi_i}, \quad (5.76)$$

此时由于

$$\pi_{ij} = \pi_i \pi_j, \quad (5.77)$$

故 \hat{Y}_{PS} 的方差有如下简单的形式:

$$V(\hat{Y}_{PS}) = \sum_{i=1}^N (1 - \pi_i) \frac{Y_i^2}{\pi_i}, \quad (5.78)$$

它的一个无偏估计是:

$$v(\hat{Y}_{PS}) = \sum_{i=1}^n (1 - \pi_i) \frac{y_i^2}{\pi_i^2}. \quad (5.79)$$

由于这里的 n 是随机的, 因此可以考虑如下的比估计:

$$\hat{Y}'_{PS} = \begin{cases} \hat{Y}_{PS} \cdot \frac{v}{n}, & \text{若 } n > 0; \\ 0, & \text{若 } n = 0. \end{cases} \quad (5.80)$$

\hat{Y}'_{PS} 是近似无偏的, 它的一个近似均方误差或方差由下式给出:

$$V(\hat{Y}'_{PS}) \approx \sum_{i=1}^N \pi_i (1 - \pi_i) \left(\frac{Y_i}{\pi_i} - \frac{Y}{\nu} \right)^2 + p_0 Y^2, \quad (5.81)$$

其中 p_0 是 $n=0$, 也即在一轮 Poisson 抽样抽到一个空样本的概率, (5.81) 式的一个估计是:

$$v(\hat{Y}'_{PS}) = \sum_{i=1}^n (1 - \pi_i) \left(\frac{y_i}{\pi_i} - \frac{\hat{Y}'_{PS}}{\nu} \right)^2 + p_0 \hat{Y}_{PS}'^2. \quad (5.82)$$

为估计 p_0 , Ogus 与 Clark(1971) 考虑了以下的修正 Poisson 抽样(modified Poisson sampling). 如果在一轮 Poisson 抽样中抽到一个空样本, 则重新进行一轮 Poisson 抽样, 也即再做 N 次 Bernoulli 试验, 直到得到一个非空样本为止. 在一轮 Poisson 抽样中, 抽到第 i 个单元的概率为 $\pi_i(1-p_0)$, 因此 p_0 满足

$$p_0 = \prod_{i=1}^N [1 - \pi_i(1 - p_0)]. \quad (5.83)$$

用迭代法即可求得 p_0 , 初始值可取为 0. 对于修正的 Poisson 抽样, 有

$$\pi_{ij} = \pi_i \pi_j (1 - p_0) \quad (i \neq j). \quad (5.84)$$

于是按(5.76)的 Horvitz-Thompson 估计, 记为 $\hat{Y}_{m_{PS}}$ 的方差公式是

$$V(\hat{Y}_{m_{PS}}) = \sum_{i=1}^N (1 - \pi_i) \frac{Y_i^2}{\pi_i} - p_0 \left(Y^2 + \sum_{i=1}^N Y_i^2 \right), \quad (5.85)$$

它的一个无偏估计是

$$v(\hat{Y}_{m_{PS}}) = \sum_{i=1}^n (1 - \pi_i) \frac{y_i^2}{\pi_i^2} - \frac{p_0}{1 - p_0} \left(\hat{Y}_{m_{PS}}^2 - \sum_{i=1}^n \frac{y_i^2}{\pi_i^2} \right). \quad (5.86)$$

若按形如(5.80)的比估计, 记为 $\hat{Y}'_{m_{PS}}$, 近似方差为

$$V(\hat{Y}'_{m_{PS}}) \approx \sum_{i=1}^N \pi_i [1 - (1 - p_0)\pi_i] \left(\frac{Y_i}{\pi_i} - \frac{Y}{\nu} \right)^2, \quad (5.87)$$

它可用下式估计:

$$v(\hat{P}'_{m_{PS}}) = \sum_{i=1}^N [1 - (1 - p_0)\pi_i] \left(\frac{y_i}{\pi_i} - \frac{\hat{P}'_{m_{PS}}}{\nu} \right)^2. \quad (5.88)$$

通常修正的 Poisson 抽样估计的方差小于一般的 Poisson 抽样估计的方差.

5.5.4 配置抽样

这是由 Brewer、Early 与 Joyce(1972) 提出的另一种严格 π PS、严格不放回, 但 n 不事先固定的抽样方法. 与 Poisson 抽样类似, 配置抽样 (collocated sampling) 先给每个单元赋予一个入样概率 π_i , 使 $\pi_i/Z_i = \nu$, 其中 ν 为一常数. 等概率地给总体单元配置一组序号 L_1, L_2, \dots, L_N ($L_i = 1, 2, \dots, N$), 在 $[0, 1]$ 中抽取一个随机数 r , 令

$$r_i = (L_i + r - 1)/N. \quad (5.89)$$

若 $r_i < \pi_i$, 则第 i 个单元入样, 否则, 该单元不入样. 对所有单元都按上述准则确定其入样与否, 构成一轮配置抽样, 因此, 实际样本量 n 是随机的.

对总体总和 Y 采用比估计型的估计量:

$$\hat{P}_{CS} = \begin{cases} \frac{\nu}{n} \sum_{i=1}^n \frac{y_i}{\pi_i}, & \text{若 } n > 0; \\ 0, & \text{若 } n = 0. \end{cases} \quad (5.90)$$

\hat{P}_{CS} 是近似无偏的, 它的近似均方误差或近似方差为:

$$V(\hat{P}_{CS}) \approx \sum_{i=1}^N \left(\frac{Y_i}{\pi_i} - \frac{Y}{\nu} \right)^2 + 2 \sum_{i=1}^N \sum_{j=1}^N (\pi_{ij} - \pi_i \pi_j) \left(\frac{Y_i}{\pi_i} - \frac{Y}{\nu} \right) \left(\frac{Y_j}{\pi_j} - \frac{Y}{\nu} \right) + P_{00} Y^2. \quad (5.91)$$

其中 P_{00} 是在一轮抽样中抽到一个空样本的概率. $V(\hat{P}_{CS})$ 的一个估计为:

$$v(\hat{P}_{CS}) = \sum_{i=1}^n (1 - \pi_i) \left(\frac{y_i}{\pi_i} - \frac{\hat{P}_{CS}}{\nu} \right)^2 + 2 \sum_{i=1}^n \sum_{j=1}^n \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{\hat{P}_{CS}}{\nu} \right) \times \left(\frac{y_j}{\pi_j} - \frac{\hat{P}_{CS}}{\nu} \right) + P_{00} \hat{P}_{CS}^2. \quad (5.92)$$

为了能计算上式, 需要给出 π_{ij} 和 P_{00} 的近似表达式. 为简单起见, 不妨约定 $\pi_1 \leq \pi_2 \leq \dots \leq \pi_N$,

$$\pi_{ij} = \frac{1}{N(N-1)} \{ [N\pi_i](N\pi_i - 1) + K_i[N\pi_i] + \max(K_j - K_i, 0) \}, \quad (5.93)$$

$$P_{oc} = \begin{cases} \frac{1}{N!} \prod_{i=1}^v (i - N\pi_i), & \text{若 } \min_i (i - N\pi_i) > 0; \\ 0, & \text{否则.} \end{cases} \quad (5.94)$$

(5.93)式中的 $K_i = N\pi_i - [N\pi_i]$, 而 $[N\pi_i]$ 表示不超过 $N\pi_i$ 的最大整数.

5.5.5 不同抽样或估计方法性质的比较

前面介绍了几种不放回的 π PS 抽样及其相应的估计量. 在实际工作中, 究竟选择哪一种方法, 要根据各种方法的特点及关于它的综合评价. 在这一小节中, 我们对前面讨论过的方法进行大致的比较. 表 5.4 是对 $n=2$ 的诸方法所作的比较, 以 Brewer 方法为标准; 表 5.5 是对 $n>2$ 的诸方法所作的比较, 以 Rao-Sampford 方法为标准. 其中效率一栏以总体总量的 Horvitz Thompson 估计 \hat{P}_{HT} 及其 Yates-Grundy-Sen 方差估计 v_{YGS} 作为标准. 在作此比较时, 我们基于抽样调查中的一个常用的理论模型——线性随机模型 (linear stochastic model). 模型的表述如下:

将所考察的总体看成是从一个无限超总体按一定随机模式产生的一个 (大小为 N) 的样本, 对每个假定的总体 (超总体的一个样本), π_i 都严格地与单元的大小成比例, 且为常数. 令 Z_i 为标准化了的大小. 模型假定为:

- 1) $Y_i = \beta Z_i + \varepsilon_i$
- 2) $\mathcal{E}(\varepsilon_i) = 0$
- 3) $\mathcal{E}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma_i^2, & J = I \\ 0, & J \neq I; \end{cases}$

表 5.4 若干 $n=2$ 的 π PS 抽样方法 (估计量) 性质的比较

方法 (估计量)	Yates-Grundy 法 \hat{P}_R	Yates-Grundy 法 \hat{P}_H	RHC 法 \hat{P}_{RHC}	Brewer 法 \hat{P}_{BT}
n 是否固定	是	是	是	是
方差估计是否无偏	是	是	是	是
方差估计的稳定性	好	很好	极好	标准
方差估计简单程度	非常简单	非常简单	很简单	简单
效率:	接近标准	高于标准	高于标准	标准
$\gamma \leq \frac{1}{2}$				
$\frac{1}{2} < \gamma < 1$	低于标准	近似为标准	低于标准	标准
$\gamma = 1$	远低于标准	低于标准	远低于标准	标准

$$4) \sigma_f^2 = \sigma^2 Z_f^2 \gamma, \quad \frac{1}{2} \leq \gamma \leq 1.$$

其中 β, σ^2, γ 皆为常数, \mathcal{E} 为超总体中的期望算子, 即对其中所有可能的总体求期望.

表 5.5 若 $n > 2$ 的 π, S 抽样方法(估计量)性质的比较

方法(估计量)	Yates-Grundy 法 \hat{Y}_G	Yates-Grundy 法 \hat{Y}_M	BHC 法 \hat{Y}_{BHC}	修正 Poisson 法 \hat{Y}_{HPS}	配置抽样 法 \hat{Y}_{ca}	Rao- Sampford 法 \hat{Y}_{RS}
n 是否固定	是	是	是	否	否	是
估计是否无偏	是	是	是	近似	近似	是
方差估计是否无偏	是	是	是	近似	近似	是
方差估计的稳定性	好	很好	极好	不知道	不知道	标准
抽样的简单程度	非常简单	相当简单	很简单	非常复杂	需计算机	简单
方差估计简单程度	非常简单	复杂	非常简单	简单	简单	简单
效率:						
$\gamma \leq 1/2$	近似标准	远高于标准	远高于标准	近似标准	近似标准	标准
$1/2 < \gamma < 1$	低于标准	近似标准	低于标准	近似标准	近似标准	标准
$\gamma = 1$	远低于标准	低于标准	远低于标准	近似标准	近似标准	标准

第 6 章

整 群 抽 样

§ 6.1 引 言

6.1.1 定 义

定义 6.1 设总体由 N 个大单元, 即初级单元 (primary unit) 组成, 每个初级单元又由若干个较小的次级单元或二级单元 (secondary unit) 组成. 从总体中按某种方式抽取 n 个初级单元, 观测其中所包含的所有次级单元. 这种抽样称为整群抽样 (cluster sampling).

确切地说, 上述抽样应称为单阶整群抽样 (single-stage cluster sampling). 如果总体中的单元可以分成多级, 则可以对前几级单元采用多阶抽样 (详见下章), 而在最后一阶中对该级抽样单元中所包含的全部最低级单元进行观测, 即是多阶整群抽样 (multi-stage cluster sampling).

在整群抽样中, 那些一旦被抽中即需观测其中所有最低级单元的单元, 例如单阶整群抽样中的初级单元, 称为整群抽样单元 (cluster sampling unit) 或简称为群 (cluster). 由于实际 (最后一阶) 抽样是整群进行的, 故称整群抽样. 本章只讨论单阶整群抽样.

6.1.2 适用场合及实施理由

整群抽样的应用颇为广泛, 其原因主要有以下几方面:

一、缺少次级单元的抽样框

在有些调查中, 尽管调查对象是较小的单元, 即上述的次级单元 (或最低级单元), 但在总体中没有或不易得到包括所有这些单元的抽样框, 也不值得为此搞一个. 例如对一个城市就很难有一份现成的包含所有居民或房屋的名册或清单. 但有可能搞到或较易编制关于较大单元 (例如居委会或户) 的抽样框, 因而可以按较大的单元进行抽样.

二、实施便利, 节省费用

即使关于次级单元的抽样框可以获得, 但从经济上考虑, 直接按次级

单元抽样获得的样本必然会相当分散。从而使调查不方便,大大增加了诸如旅费之类的费用,耗时也更多。相反的,按整群抽样,由于样本相对集中,调查既方便,费用也节省。例如在某城市开展家用电器调查,若调查的最低单元是户。在全市中抽取300户进行调查比抽取15个居民小组(设平均每个居民小组包含20户)进行调查所费的时间与经费要多得多。又例如在人体尺寸调查中,每个被测的人要测量50多项指标,且必须在专业人员指导下用整套专用测量仪器来测量。在此情形,以数十人为一群体进行抽样(相当于一天的工作量)显然要比以个人为抽样单元方便得多。虽然对同样数目的小单元而言,整群抽样的精度可能有所损失,但因每调查一个小单元的平均费用(或耗时)低,故可以通过适当增大样本量的方法来得到弥补。例如前面提到的家电调查,抽30个居民小组共600户进行调查,其结果可能比用简单随机抽样在全市调查300户的做法既省费用且精度也高。

三、对某些特殊结构的总体,有较高的精度

例如为估计一个地区的男女性别的比例,由于每个家庭内成员的性别结构有一定的模式,此时对户采用整群抽样的精度比直接抽人的精度高得多(参见例6.2与6.4)。

6.1.3 群划分的原则

关于群的划分,有两个问题:一是如何定义群,即当群并非是一个自然形成的单位时,确定每个群的组成;二是如何确定群的规模即群的大小。

对于前一个问题,群的划分应尽可能使群与群之间的差异小,而群内差异则愈大愈好。这样,每个群都具有足够好的代表性。如果所有的群都相似,那么抽少数群就可获得相当好的精度;反之,若群内的单元比较相似,而群与群之间的差别较大,则整群抽样的效率就低。所以分群的原则与分层的原则是恰好相反的。图6.1直观地表明了理想的分群与分层的思想,其中同一字母表示有相近的观测值的单元。

至于群的规模的选择,一是取决于精度与费用之间的平衡,二是从抽样实施的组织管理等因素来考虑。对于前者,群的规模选得大,则费用省而精度差;群的规模选得小,则精度高而费用大。这方面除了依靠实践经验外,还可对假定的方差函数与费用函数作理论上的最优选择。

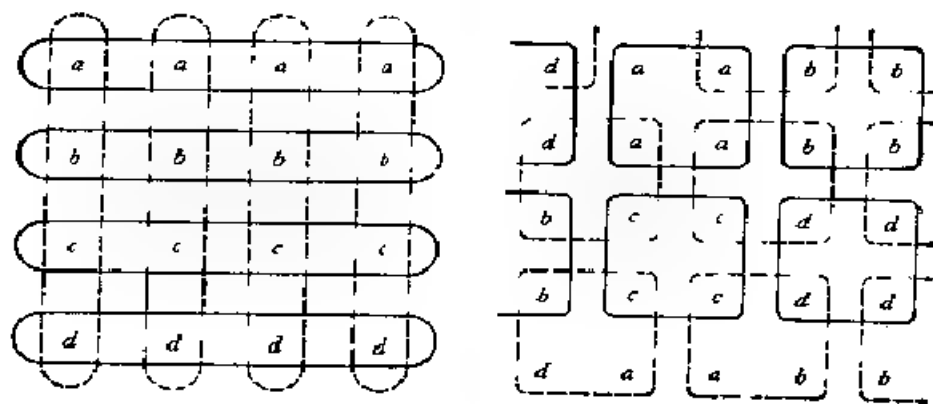


图 6.1 理想的群(虚线)与层(实线)的划分

§ 6.2 群大小相等的情形

本节首先讨论群的规模, 即群的大小或它所包含的次级单元个数都相等的情形. 假定对群的抽样是简单随机的.

6.2.1 记 号

记 Y_{ij} 为第 i 群(初级单元)中第 j 个次级单元的观测值($i=1, 2, \dots, N; j=1, 2, \dots, M$, 其中 M 是群的大小).

y_{ij} 是样本中第 i 初级单元中第 j 次级单元的观测值($i=1, 2, \dots, n; j=1, 2, \dots, M$).

$$Y_i = \sum_{j=1}^M Y_{ij},$$

$$\bar{Y}_i = Y_i / M,$$

$$\bar{Y} = \sum_{i=1}^N Y_i / N,$$

$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M Y_{ij} / NM = Y / M,$$

$$S^2 = \frac{1}{NM-1} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y})^2,$$

$$S_b^2 = \frac{M}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2,$$

$$S_u^2 = \frac{1}{N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2,$$

$$f = \frac{n}{N}.$$

$$y_i = \sum_{j=1}^M y_{ij},$$

$$\bar{y}_i = y_i / M,$$

$$\bar{y} = \sum_{i=1}^n y_i / n,$$

$$\bar{y} = \sum_{i=1}^n \sum_{j=1}^M y_{ij} / nM = \bar{Y} / M,$$

$$s^2 = \frac{1}{nM-1} \sum_{i=1}^n \sum_{j=1}^M (y_{ij} - \bar{y})^2,$$

$$s_b^2 = \frac{M}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2,$$

$$s_u^2 = \frac{1}{n(M-1)} \sum_{i=1}^n \sum_{j=1}^M (y_{ij} - \bar{y}_i)^2,$$

6.2.2 总体与样本平方和的分解

正如上节中所述的, 整群抽样的精度在很大程度上取决于群内次级单元差异的大小, 或者说取决于群内次级单元相似程度的大小, 为此, 运用方差分析的方法, 将总体与样本中所有单元的观测值对总体(按次级单元)均值 \bar{Y} 或样本均值 \bar{y} 的(离差)平方和进行分解是有用的。

对于总体, Y_{ij} 对 \bar{Y} 离差的总平方和可以分解为:

$$\sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^N \sum_{j=1}^M [(Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y})]^2 \\ = \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2 + M \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2, \quad (6.1)$$

其中第一项是群内平方和:

$$\sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2 = N(M-1)S_u^2, \quad (6.2)$$

$N(M-1)$ 是它的自由度, 而 S_u^2 即是群内方差。同样, (6.1) 中的第二项

$$M \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2 = (N-1)S_b^2 \quad (6.3)$$

是群间的平方和, 它的自由度为 $N-1$, S_b^2 是群间的方差。

根据总体方差 S^2 的定义及上述平方和的分解, 我们有:

$$S^2 = \frac{1}{NM-1} [(N-1)S_b^2 + N(M-1)S_u^2]. \quad (6.4)$$

上述结果可写成熟知的方差分析表 6.1。

表 6.1 Y_{ij} 的方差分析

变 差 来 源	自 由 度	平 方 和	方 差
群 间	$N-1$	$M \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2$	S_b^2
群 内	$N(M-1)$	$\sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2$	S_u^2
总 计	$NM-1$	$\sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y})^2$	S^2

完全类似的, 对样本观测值也可作同样的分解, 相应的方差分析表如表 6.2 所示。

其中样本方差 s^2 与样本群间的方差 s_b^2 与群内方差 s_u^2 的关系有:

$$s^2 = \frac{1}{nM-1} [(n-1)s_b^2 + n(M-1)s_u^2]. \quad (6.5)$$

表 6.2 y_{ij} 的方差分析

变 差 来 源	自 由 度	平 方 和	方 差
群 间	$n-1$	$M \sum_i (\bar{y}_i - \bar{y})^2$	s_b^2
群 内	$n(M-1)$	$\sum_i \sum_j (y_{ij} - \bar{y}_i)^2$	s_w^2
总 计	$nM-1$	$\sum_i \sum_j (y_{ij} - \bar{y})^2$	s^2

注意: 此时 s^2 并不是 S^2 的无偏估计, 这是因为按次级单元而言, 样本并不是简单随机的, 但由于对群的抽取是简单随机的, 因此, s_b^2 与 s_w^2 分别是 S_b^2 与 S_w^2 的无偏估计. 为了证明这一点, 只要注意 s_b^2/M 是 \bar{y}_i 的样本方差, 它是相应的总体方差 S_b^2/M 的无偏估计, 而 s_w^2 则可以看作是 $z_i \triangleq \sum_{j=1}^M (y_{ij} - \bar{y}_i)^2 / (M-1)$ 的样本均值, 于是它是 Z_i 的总体均值 \bar{Z} 也即 S_w^2 的无偏估计.

根据上述结论及(6.4)式, S^2 的一个无偏估计可以构造如下:

$$\hat{S}^2 = \frac{(N-1)s_b^2 + N(M-1)s_w^2}{NM-1}. \quad (6.6)$$

当 N 很大时,

$$\hat{S}^2 \approx \frac{s_b^2 + (M-1)s_w^2}{M}. \quad (6.7)$$

另一方面, 若 n 也足够大, 则 s^2 也近似地可表为(6.7)式. 因此, 只有在此时, s^2 可看作是 S^2 的近似无偏估计.

6.2.3 群内相关 ρ_c

定义 6.2 同一群内不同次级单元的观测值对总体均值离差乘积的平均与总体所有次级单元观测值对总体均值离差平方的平均之比 ρ_c 称为群内相关(intraclass correlation coefficient), 即下式:

$$\rho_c = \frac{E(Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y})}{E(Y_{ij} - \bar{Y})^2} \quad (6.8)$$

$$= \frac{2 \sum_{i=1}^N \sum_{j \neq k}^M (Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y})}{(M-1)(NM-1)S^2}. \quad (6.9)$$

其中(6.8)式的分子是 $N \binom{M}{2} = NM(M-1)/2$ 个次级单元对 Y_{ij} 、 Y_{ik}

($j < k$) 对 \bar{Y} 的离差乘积和的平均:

$$E(Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y}) = \frac{\sum_{i=1}^N \sum_{j < k}^M (Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y})}{NM(M-1)/2}$$

而(6.8)式的分母是 MN 个 Y_{ij} 对 \bar{Y} 的离差平方和的平均:

$$E(Y_{ij} - \bar{Y})^2 = \frac{\sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y})^2}{MN} = \frac{NM-1}{MN} S^2.$$

因此(6.8)与(6.9)两式相等.

ρ_c 的取值在 $\left[-\frac{1}{M-1}, 1\right]$ 范围内. 当 $\rho_c = 0$ 时, 表明群完全是随机组成的. ρ_c 值愈大, 表明群内的单元愈相似. ρ_c 值愈小, 则群内单元的差异愈大. 当 $\rho_c < 0$ 时, 表明这个差异比随机分组时群内的差异更大.

ρ_c 可以用群间方差 S_b^2 与群内方差 S_w^2 来表示. 考虑 Y_i 对 \bar{Y} 的离差平方和:

$$\begin{aligned} \sum_{i=1}^N (Y_i - \bar{Y})^2 &= \sum_{j=1}^N M \bar{Y}_j - M \bar{Y})^2 \\ &= M \sum_{j=1}^N M (\bar{Y}_j - \bar{Y})^2 = M(N-1) S_b^2, \end{aligned}$$

同时它又可表成:

$$\begin{aligned} \sum_{i=1}^N (Y_i - \bar{Y})^2 &= \sum_{i=1}^N \left[\sum_{j=1}^M (Y_{ij} - \bar{Y}) \right]^2 \\ &= \sum_{i=1}^N \left[\sum_{j=1}^M (Y_{ij} - \bar{Y})^2 + 2 \sum_{j < k} (Y_{ij} - \bar{Y})(Y_{ik} - \bar{Y}) \right] \\ &= (NM-1) S^2 + (NM-1)(M-1) S^2 \rho_c \\ &= (NM-1) S^2 [1 + (M-1) \rho_c]. \end{aligned}$$

因而

$$1 + (M-1) \rho_c = \frac{M(N-1) S_b^2}{(NM-1) S^2}$$

故

$$\rho_c = \frac{M(N-1) S_b^2 - (NM-1) S^2}{(M-1)(NM-1) S^2} \approx \frac{S_b^2 - S^2}{(M-1) S^2}. \quad (6.10)$$

另一方面,

$$\begin{aligned} (NM-1) S^2 &= \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y})^2 \\ &= \sum_{i=1}^N M (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2 \\ &= \frac{1}{M} \sum_{i=1}^N (Y_i - \bar{Y})^2 + N(M-1) S_w^2 \end{aligned}$$

$$= \frac{1}{M} (NM - 1) S^2 [1 + (M - 1) \rho_c] + N(M - 1) S_w^2,$$

$$S_u^2 = \frac{(NM - 1)(1 - \rho_c) S^2}{MN} \approx S^2(1 - \rho_c)$$

从而

$$\rho_c = 1 - \frac{NMS_u^2}{(NM - 1)S^2} \approx 1 - \frac{S_w^2}{S^2}. \quad (6.11)$$

为估计 ρ_c , 从(6.10)式或(6.11)式出发, 注意到 S^2 可用 \hat{S}^2 估计, 因而有

$$\hat{\rho}_c \approx \frac{s_b^2 - s_w^2}{s_b^2 + (M - 1)s_w^2}. \quad (6.12)$$

在实际问题中, 当群的大小 M_i 不等时, 上述公式也能适用. 此时按通常的平方和分解方法计算 s_b^2 与 s_w^2 , 用平均群的大小 \bar{M} 代替 M 即可.

例 6.1 在一次对居民月收入的试调查中, 按简单随机抽样抽得 $n = 10$ 个居民小组, 各居民小组的平均户月收入 \bar{y}_i 及标准差 s_i 如表 6.3 所示. 平均每个居民小组包含 $M = 16$ 户, 求群内相关 ρ_c .

表 6.3 10 个居民小组的户平均月收入及标准差

i	y_i	s_i	i	\bar{y}_i	s_i
1	758.9	78.2	6	796.1	102.4
2	766.0	102.7	7	720.5	93.3
3	821.3	184.6	8	812.4	145.5
4	825.5	121.5	9	733.8	72.0
5	689.4	70.9	10	773.7	89.1

解 根据表 6.3 中的数据可计算:

$$\bar{y} = \frac{1}{10} \sum_{i=1}^{10} \bar{y}_i = 782.76,$$

$$s_b^2 = \frac{16}{9} \sum_{i=1}^{10} (\bar{y}_i - \bar{y})^2 = 4210.30 \times 16 = 67364.8,$$

$$s_w^2 = \frac{1}{n} \sum_{i=1}^{10} \frac{1}{M-1} \sum_{j=1}^M (y_{ij} - \bar{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{10} s_i^2 = 12401.2,$$

$$\therefore \hat{\rho}_c = \frac{67364.8 - 12401.2}{67364.8 + (16 - 1) \times 12401.2} = 0.217.$$

例 6.2 为估计某街道男性居民在全体居民中的比例, 按家庭户进行整群抽样, 共抽得 500 个家庭, 按家庭人口数及性别结构分类, 分类及相应的频数见表 6.4.

表 6.4 500 个家庭户人口性别结构分类情况

类别 k	家庭人口数 m_k	男性人数 b_k	女性人数 $m_k - b_k$	频数 n_k
1	1	1	0	3
2	2	0	2	1
3	2	1	1	34
4	2	2	0	2
5	3	1	2	96
6	3	2	1	99
7	4	1	3	51
8	4	2	2	94
9	4	3	1	47
10	5	1	4	10
11	5	2	3	19
12	5	3	2	21
13	5	4	1	11
14	6	2	4	2
15	6	3	3	7
16	6	4	2	3

根据表 6.4, 在 500 个样本户中:

总人口数 $\sum_k n_k m_k = 1807,$

平均每户人口数 $\bar{m} = \frac{1807}{500} = 3.614,$

男性人口数 $\sum_k n_k b_k = 907,$

男性比例 $p = \frac{907}{1807} = 0.501937,$

女性人口数 $1807 - 907 = 900,$

女性比例 $1 - p = 0.498063,$

总平方和 $(\sum_k n_k m_k) p(1-p) = 451.7432,$

群(家庭)内平方和 $\sum_k n_k \frac{b_k(m_k - b_k)}{m_k} = 396.4667,$

群内平方和的自由度 $\sum_k n_k (m_k - 1) = 1307,$

群(家庭)间平方和 $451.7432 - 396.4667 = 55.2765,$

群间平方和自由度 $1807 - 1 - 1307 = 499,$

$$s_b^2 = \frac{55.2765}{499} = 0.1108,$$

$$s_{10}^2 = \frac{396 \cdot 4667}{1307} = 0.3033,$$

$$\therefore \hat{\rho}_c = \frac{s_b^2 - s_w^2}{s_b^2 + (\bar{m} - 1)s_w^2} = \frac{0.1108 - 0.3033}{0.1108 + 2 \cdot 614 \times 0.3033} = -0.213.$$

6.2.4 估计量及其方差

定理 6.1 对整群抽样, 若群的抽取是简单随机的, 且群的大小皆等于 M , 则

$$\bar{y} = \sum_{i=1}^n \sum_{j=1}^M y_{ij} / nM \quad (6.13)$$

是总体均值

$$\bar{Y} = \sum_{i=1}^N \sum_{j=1}^M Y_{ij} / NM$$

的无偏估计. 又 \bar{y} 的方差为:

$$V(\bar{y}) = \frac{1-f}{n} \cdot \frac{NM}{M^2(N-1)} S^2 [1 + (M-1)\rho_c] \quad (6.14)$$

$$\approx \frac{1-f}{nM} S^2 [1 + (M-1)\rho_c], \quad (6.15)$$

其中 $f = n/N$.

证明 由于群是按简单随机方法抽取的, 因此 $y = \frac{1}{n} \sum_{i=1}^n y_i = M\bar{y}$ 是 $\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i = M\bar{Y}$ 的无偏估计, 因而 \bar{y} 是 \bar{Y} 的无偏估计. 又

$$\begin{aligned} M^2 V(\bar{y}) &= V(y) = \frac{1-f}{n} \cdot \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N-1}, \\ \therefore V(\bar{y}) &= \frac{1-f}{nM^2} \cdot \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N-1} \\ &= \frac{1-f}{nM^2} \cdot \frac{NM(N-1)S^2 [1 + (M-1)\rho_c]}{N-1} \\ &\approx \frac{1-f}{nM} S^2 [1 + (M-1)\rho_c]. \quad \blacksquare \end{aligned}$$

推论 $V(\bar{y})$ 的一个无偏估计为:

$$v(\bar{y}) = \frac{1-f}{n} \cdot \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{1-f}{nM} s_b^2. \quad (6.16)$$

6.2.5 设计效应

根据定理 6.1, 容易得到整群抽样的设计效应. 如果在总体中按次

级单元抽取样本量为 nM 的简单随机样本, 则

$$V_{\text{ran}}(\bar{y}) = \frac{1-f}{nM} S^2$$

与(6.15)式比较, 即可得到整群抽样的设计效应

$$\text{deff} = \frac{V(\bar{y})}{V_{\text{ran}}(\bar{y})} \approx 1 + (M-1)\rho_c, \quad (6.17)$$

在实际问题中, 若群大小 M_i 不完全相等, 则可用平均群的大小 \bar{M} 代替 M .

由于一般的 $\rho_c > 0$, 因此从(6.17)知, 整群抽样的精度在大多数情况下, 比抽同样数量的次级单元的简单随机抽样的精度低. 为了获得与简单随机抽样相同的精度, 则整群抽样的样本量必须是简单随机抽样样本量的 $1 + (M-1)\rho_c$ 倍. 通常整群抽样的样本量即是根据此确定的.

例 6.3 在对全国成年人人体尺寸测量中, 根据一次试测样本的分析, 单位内同性别人的群内相关的估计为 $\hat{\rho}_c = 0.00775$. 根据精度要求, 按简单随机抽样所需的样本量为 $n_0 = 6147$. 若平均群的大小为 $\bar{M} = 80$, 则按单位的整群抽样

$$\text{deff} = 1 + (80-1) \times 0.00775 = 1.61225.$$

从而 $n = n_0 \cdot \text{deff} = 6147 \times 1.61225 = 9911$.

也即需抽 9911 人, 合 124 个群.

例 6.4 (续例 6.2) 为估计男性居民在全体居民中的比例, 用整群抽样抽取 500 户共 1807 人. 根据例 6.2 中的计算 $\hat{\rho}_c = -0.213$, 因而在此问题中, 按户整群抽样的设计效应:

$$\text{deff} = 1 + (\bar{M}-1)\hat{\rho}_c = 0.4432.$$

在这个特殊问题中, 整群抽样的效果反而比简单随机抽样高. 这是因为在一个家庭内由夫妻为核心加上其子女或父母, 本身就存在一定的性别结构, 因此家庭内(群内)性别的差异必然比随机分组产生的组内差异大. 在此例中, 要达到整群抽样实际精度的简单随机抽样, 需抽取

$$\frac{1807}{\text{deff}} = \frac{1807}{0.4432} = 4077(\text{人}).$$

§ 6.3 对比例估计的整群抽样

本节考虑用整群抽样来估计具有某种特征的(次级)单元在总体中所占的比例 P . 在实际问题中, 对 P 的估计常用整群抽样, 因为它不仅方

便, 而且对某些特殊问题(如例 6.4), 精度也高. 因此对比例估计采用整群抽样, 总的效率是高的. 本节仍考虑群的抽取是简单随机的.

6.3.1 群大小相等的情形

如果总体中的群的大小都相等(或近似相等), 则可直接利用第二章中简单随机抽样的结果. 令 a_i 为第 i 个群(初级单元)中具有所考虑特征的次级单元数(相当于上节中的 y_i), 令 $p_i = \frac{a_i}{M}$ 是样本中第 i 群中具有所考虑特征的次级单元的比例, 则有如下定理:

定理 6.2 按简单随机抽样在 N 个初级单元(群)中抽取 n 个, 则

$$p = \frac{1}{n} \sum_{i=1}^n p_i = \frac{1}{nM} \sum_{i=1}^n a_i \quad (6.18)$$

是总体中具有所考虑特征的次级单元的比例 P 的无偏估计, 且

$$V(p) = \frac{1-f}{n} \frac{\sum_{i=1}^N (P_i - P)^2}{N-1}, \quad (6.19)$$

又

$$v(p) = \frac{1-f}{n} \frac{\sum_{i=1}^n (p_i - p)^2}{n-1} \quad (6.20)$$

是 $V(p)$ 的无偏估计.

证明 将 $p_i (P_i)$ 作为(初级)单元的指标, p 是样本均值, P 是总体均值, 于是由定理 2.1、定理 2.2 及定理 2.4 的推论即获得相应的结论.

根据(6.19)式即可计算比例估计整群抽样的设计效应. 若对次级单元直接进行简单随机抽样, 抽取 nM 个次级单元, 则

$$V_{\text{ran}}(p) = \frac{NM - nM}{NM - 1} \frac{PQ}{nM} \approx \frac{1-f}{n} \cdot \frac{PQ}{M}.$$

因而整群抽样的设计效应

$$\text{deff} = \frac{V(p)}{V_{\text{ran}}(p)} \approx \frac{M \sum_{i=1}^N (P_i - P)^2}{NPQ}. \quad (6.21)$$

这里 $Q = 1 - P$. 如果每个 P_i 与 P 差别不大, 则整群抽样的效率就比较高.

6.3.2 群大小不相等的情形——比估计

对于比例估计, 在群大小 M_i 不相等时, 用比估计方法很容易处理.

此时 P 的一个自然的估计是:

$$p = \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n m_i}. \quad (6.22)$$

其中 m_i 是抽中的第 i 个群的大小. 由于 P 实际上是样本两个总和之比, 因此根据定理 4.1, 有如下定理:

定理 6.3 若群的抽取是简单随机的, 则对总体比例 P 的估计 p , 当 n 大时是近似无偏的, 且

$$V(p) \approx \frac{1-f}{n\bar{M}^2} \cdot \frac{\sum_{i=1}^N (a_i - PM_i)^2}{N-1} = \frac{1-f}{n\bar{M}^2} \cdot \frac{\sum_{i=1}^N M_i^2 (P_i - P)^2}{N-1}, \quad (6.23)$$

其中 $\bar{M} = \frac{1}{N} \sum_{i=1}^N M_i$ 是总体群的平均大小. 又 $V(p)$ 可用下式估计:

$$v(p) = \frac{1-f}{n\bar{m}^2} \cdot \frac{1}{n-1} \left(\sum_{i=1}^n a_i^2 + p^2 \sum_{i=1}^n m_i^2 - 2p \sum_{i=1}^n a_i m_i \right). \quad (6.24)$$

证明 在定理 4.1 及其推论中, 用 p 代替 $\hat{\theta}$, 用 m_i 代替 ω_i , 用 a_i 代替 y_i , 即可获证. ■

例 6.5 (续例 6.2、例 6.4) 随机抽取 500 个家庭, 估计男性居民的比例 P . 根据表 6.4 中的数据:

$$n = 500, \quad \sum_{i=1}^n a_i = 907, \quad \sum_{i=1}^n m_i = 1807, \quad \bar{m} = 3.614, \quad p = \frac{907}{1807} = 0.501937,$$

$$\begin{aligned} v(p) &= \frac{1}{n\bar{m}^2} \cdot \frac{1}{n-1} \left(\sum_{i=1}^n a_i^2 + p^2 \sum_{i=1}^n m_i^2 - 2p \sum_{i=1}^n a_i m_i \right) \\ &= \frac{1}{n\bar{m}^2(n-1)} \left(\sum_k n_k a_k^2 + p^2 \sum_k n_k m_k^2 - 2p \sum_k n_k a_k m_k \right) \\ &= \frac{1}{500 \times (3.614)^2 \times 499} \left[1957 + \left(\frac{907}{1807} \right)^2 \times 6935 - 2 \left(\frac{907}{1807} \right) \right. \\ &\quad \left. \times 3478 \right] = 6.528189 \times 10^{-5}, \end{aligned}$$

$$s(p) = \sqrt{v(p)} = 0.00808.$$

在例 6.4 中已计算为得到整群抽样相同精度的简单随机抽样的样本量应为 $n = 4077$ 人. 作为验证, 此时按简单随机抽样对比例估计的标准差估计(取 $p = 0.5$):

$$s'(p) = \sqrt{v_{\text{ran}}} = \sqrt{\frac{pq}{n-1}} = 0.00783,$$

这与前面的结果相吻合(由于例 6.4 中对 deff 的估计是按等群大小计算的, 因此稍有一些误差).

§ 6.4 群大小不等的一般情形

在大多数情形, 群大小 M_i 是不相等的. 此时, 若 M_i 相差不多, 则仍可按 § 6.2 中的方法处理, 用平均群大小 $\bar{M} = \sum_{i=1}^N M_i / N$ 代替 M . 或先根据群的大小分层, 在层内按上面的方法处理. 当需要估计比例时, 则可用 6.3.2 段的比估计方法处理. 但是对群大小不相等的一般情形, 若仍对群进行简单随机抽样, 并取简单估计, 则一般的效果欠佳. 此时需对估计量进行改进或改变抽样方法, 对群进行不等概率抽样.

我们将 6.2.1 段中的记号作相应的改变.

记 Y_{ij} 为第 i 个群中第 j 个次级单元的观测值 ($i = 1, 2, \dots, N$; $j = 1, 2, \dots, M_i$, 其中 M_i 是群的大小).

y_{ij} 为样本中第 i 个群中第 j 个次级单元的观测值 ($i = 1, 2, \dots, n$; $j = 1, 2, \dots, m_i$, m_i 是群的大小).

$M_0 = \sum_{i=1}^N M_i$ 是总体中的次级单元总数.

$$\begin{aligned} Y_i &= \sum_{j=1}^{M_i} Y_{ij}, & y_i &= \sum_{j=1}^{m_i} y_{ij}, \\ \bar{Y}_i &= \sum_{j=1}^{M_i} Y_{ij} / M_i, & \bar{y}_i &= \sum_{j=1}^{m_i} y_{ij} / m_i, \\ \bar{Y} &= \sum_{i=1}^N Y_i / N, & \bar{y} &= \sum_{i=1}^n y_i / n, \\ \bar{\bar{Y}} &= \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij} / M_0, & \bar{\bar{y}} &= \sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij} / m_0. \end{aligned}$$

注意此时 $\bar{\bar{y}} \neq \sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij} / \sum_{i=1}^n m_i$.

为便于讨论及简化表达式起见, 在本节中主要讨论对总体总和 $Y = \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij}$ 的估计, 对总体平均数 \bar{Y} 的估计可以从对 Y 的估计推出来.

6.4.1 按简单随机抽样抽群——简单估计

若对群的抽样是按简单随机抽样抽取的, 将每个群和 Y_i 看作为第 i

个群的指标, 则根据定理 2.1, 立即可以得到总体总和 $Y = \sum_{i=1}^N Y_i$ 的简单估计:

$$\hat{P} = \frac{N}{n} \sum_{i=1}^n y_i = N\bar{y}. \quad (6.25)$$

\hat{P} 是无偏的, 它的方差为:

$$V(\hat{P}) = \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{N-1}. \quad (6.26)$$

其中 $f = \frac{n}{N}$, 而它的一个无偏估计为:

$$v(\hat{P}) = \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}. \quad (6.27)$$

从(6.25)可得到 \bar{Y} 的简单估计为:

$$\hat{\bar{Y}} = \hat{P}/M_0 = \frac{N\bar{y}}{M_0} = \frac{\bar{y}}{\bar{M}}. \quad (6.28)$$

其中 $\bar{M} = \frac{M_0}{N}$

是总体群的平均大小.

从 \hat{P} 的方差公式可以看出, 它主要取决于每个群和 Y_i 的波动程度.

6.4.2 按简单随机抽样抽群——比估计

在对群进行简单随机抽样的情形, 另一种可用的估计是以群的大小 M_i 为辅助变量的比估计, 即采用

$$\hat{P}_R = M_0 \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}. \quad (6.29)$$

这里 m_i 即是第 4 章中的 x_i , M_0 即是它的总体和 X , 而总体比值 R 在这里即是 $Y/M_0 = \bar{Y}$, 于是 \bar{Y} 的估计为

$$\hat{\bar{Y}}_R = \frac{\hat{P}_R}{M_0} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}. \quad (6.30)$$

上述估计量称为对大小的比估计 (ratio-to size estimator).

根据定理 4.1, \hat{P}_R (及 $\hat{\bar{Y}}_R$) 是有偏的, 但当 n 大时, 它们是近似无偏的, 此时 \hat{P}_R 的近似方差及其估计分别为:

$$\begin{aligned}
 V(\hat{P}_R) &\approx \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^N (Y_i - \bar{Y} M_i)^2}{N-1} \\
 &= \frac{N^2(1-f)}{n} \cdot \frac{\sum_{i=1}^N M_i^2 (\bar{Y}_i - \bar{Y})^2}{N-1}, \quad (6.31)
 \end{aligned}$$

$$\begin{aligned}
 v(\hat{P}_R) &= \frac{N^2(1-f)}{n} \cdot \frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 + \left(\frac{\hat{P}_R}{M_0} \right)^2 \sum_{i=1}^n m_i^2 - 2 \left(\frac{\hat{P}_R}{M_0} \right) \sum_{i=1}^n m_i y_i \right] \\
 &= \frac{N^2(1-f)}{n} \cdot \frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 + \left(\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i} \right)^2 \sum_{i=1}^n m_i^2 - 2 \left(\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i} \right) \sum_{i=1}^n m_i y_i \right]. \quad (6.32)
 \end{aligned}$$

注意: \hat{P}_R 的方差主要取决于 \bar{y}_i 与 \bar{Y} 的差异的大小. 在多数实际情形, \bar{Y}_i 的差别不是很大. 但由于 M_i 可能变化很大, 所以 Y_i 的差别也可能很大. 因此尽管 \hat{P}_R 是有偏的, 但在大多数情形, 它的均方误差却比 \hat{P} 可能小很多. 只有当 Y_i 与 M_i 无关时, 用 \hat{P} , 效果才比较好, 但这种情况在实际问题中是不多的.

例 6.5 从共有 790 个单位的某系统中按简单随机抽样抽取 20 个单位, 关于这些单位的职工人数 m_i 、月奖金总额 y_i 及人平均月奖金 \bar{y}_i 列于表 6.5. 试估计该系统人平均月奖金 \bar{Y} . 已知该系统共有职工人数 $M_0 = 337208$ 人.

简单估计:

$$\begin{aligned}
 \hat{P} &= \frac{N}{n} \sum_{i=1}^n y_i = 790 \times \frac{1078566}{20} = 42603357 (\text{元}), \\
 \hat{\bar{Y}} &= \frac{\hat{P}}{M_0} = \frac{42603357}{337208} = 126.94 (\text{元}), \\
 v(\hat{\bar{Y}}) &= \frac{N^2}{n M_0^2} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = 1062.19, \\
 s(\hat{\bar{Y}}) &= \sqrt{v(\hat{\bar{Y}})} = 32.59 (\text{元}),
 \end{aligned}$$

对大小的比估计:

$$\begin{aligned}
 \hat{\bar{Y}}_R &= \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n M_i} = \frac{1078566}{10219} = 105.5452 (\text{元}), \\
 \hat{P}_R &= \hat{\bar{Y}}_R \cdot M_0 = 35590673 (\text{元}).
 \end{aligned}$$

表 6.5 某系统按简单随机抽样抽得的 20 个单位的职工
人数、月奖金总额与平均数的数据

样本单位号 i	职工人数 m_i	月奖金总额 y_i	人均月奖金 \bar{y}_i
1	186	20088	108
2	497	48209	97
3	78	9360	120
4	1218	141288	116
5	254	23622	93
6	330	36300	110
7	118	10020	85
8	570	47310	83
9	323	34561	107
10	45	4140	92
11	472	53808	114
12	2260	239560	106
13	398	50546	127
14	52	4472	86
15	1803	171285	95
16	368	41952	114
17	781	92158	118
18	45	5940	132
19	190	16910	89
20	231	27027	117

$$v(\hat{\bar{Y}}_R) = \frac{N^2}{n(n-1)M_0^2} \left[\sum_{i=1}^n y_i^2 + \bar{Y}^2 \sum_{i=1}^n m_i^2 - 2\bar{Y} \sum_{i=1}^n m_i y_i \right] = 13.5859,$$

$$S(\hat{\bar{Y}}_R) = \sqrt{v(\hat{\bar{Y}}_R)} = 3.69(\text{元}).$$

比较两个估计量的标准差, 可知对大小的比估计 $\hat{\bar{Y}}_R$ 远比简单估计精确得多。

6.4.3 对群进行不等概率抽样

在群大小不等的整群中, 最常用且最有效的方法是对群进行与其大小成比例的不等概率抽样。此时可用上章介绍的放回 PPS 抽样或任何一种不放回的 π PPS 抽样。在估计时, 只要将群和 Y_i 看成是它的指标, 则可直接应用 Hansen-Hurwitz 估计量或 Horvitz-Thompson 估计量。

1) 若群的抽样是按与 M_i 成比例的概率放回 PPS 抽样, 即每次抽样是按

$$Z_i = \frac{M_i}{M_0} \quad (i=1, 2, \dots, N)$$

的概率在总体中抽取第 i 个群(初级单元), 独立放回地抽取 n 个群, 其大小及观测的群和分别为 m_i 及 y_i , 则总体总和 Y 的估计为:

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{z_i} = \frac{M_0}{n} \sum_{i=1}^n \frac{y_i}{m_i} = M_0 \bar{y}, \quad (6.33)$$

其中

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{m_i} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i. \quad (6.34)$$

根据定理 5.1, \hat{Y}_{HH} 是 Y 的无偏估计, 它的方差为

$$V(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^n Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 = \frac{M_0}{n} \sum_{i=1}^n M_i (\bar{Y}_i - \bar{Y})^2, \quad (6.35)$$

它的一个无偏估计为

$$v(\hat{Y}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{y_i}{z_i} - \hat{Y}_{HH} \right)^2 = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2. \quad (6.36)$$

若估计的目标量是 \bar{Y} , 则有以下简单的形式:

$$\hat{\bar{Y}}_{HH} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i = \bar{y}. \quad (6.37)$$

它是 \bar{Y} 的无偏估计, 它的方差与方差估计分别是:

$$V(\hat{\bar{Y}}_{HH}) = \frac{1}{nM_0} \sum_{i=1}^n M_i (\bar{Y}_i - \bar{Y})^2, \quad (6.38)$$

$$v(\hat{\bar{Y}}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2. \quad (6.39)$$

与简单随机抽样中的对大小比估计 $\hat{\bar{Y}}_R$ 的情况类似, $\hat{\bar{Y}}_{HH}$ 的方差取决于 $\bar{Y}_i (\bar{y}_i)$ 的差异. 因此对于次级单元比较均匀的通常情况, 用 PPS 抽样效果很好.

2) 若群的抽样是用任何一种严格的 π PS 抽样方法时, Y 的估计应用 Horvitz-Thompson 估计:

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}. \quad (6.40)$$

它也是无偏的, 其方差与方差估计由定理 5.2 及定理 5.3 给出. 例如若用 Brewer 或 Durbin 方法抽取 $n=2$ 个群, 记样本群的编号为 1, 2, 则

$$\hat{Y}_B = \frac{Y_1}{\pi_1} + \frac{Y_2}{\pi_2} = \frac{1}{2} \left(\frac{y_1}{z_1} + \frac{y_2}{z_2} \right),$$

$$v(\hat{Y}_B) = \frac{\pi_1 \pi_2 - \pi_{12}}{\pi_{12}} \left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2} \right)^2.$$

其中

$$\pi_i = 2z_i = \frac{2m_i}{M_0} \quad (i=1, 2),$$

$$\sigma_{12} = \frac{4z_1 z_2 (1 - z_1 - z_2)}{(1 - 2z_1)(1 - 2z_2) \left[1 + \sum_{i=1}^N \frac{Z_i}{1 - 2Z_i} \right]}.$$

3) 若用其他不放回不等概率方法抽取群, 则估计时需用相应的特殊统计量, 例如用 Rao Hartley Cochran 方法抽取 n 个群, 则

$$\hat{Y}_{RHC} = \sum_{g=1}^n Z_g \frac{y_g}{z_g}.$$

其中 Z_g^* 是将总体随机划分成的第 g 个群组 (由初级单元组成的组) 的 Z_i 的和, \hat{Y}_{RHC} 也是无偏的, 它的方差估计量为:

$$v(\hat{Y}_{RHC}) = \frac{\sum_{g=1}^n N_g^2}{N^2 - \sum_{g=1}^n N_g^2} \frac{N}{\sum_{g=1}^n Z_g^*} \sum_{g=1}^n Z_g^* \left(\frac{y_g}{z_g} - \hat{Y}_{RHC} \right)^2.$$

6.4.4 数值例子——对交通运输量的调查

例 6.6 某地交通部门所属的 48 个单位的每个单位的营业性货车的标识吨位和 M_i 如表 6.6 所示. 共有总吨位 $M_0 = 11861$ t (吨). 为统计该部门某月完成的货运周转量 $Y^{(1)}$ 与运量 $Y^{(2)}$, 以 M_i 为单位大小进行放回 PPS 抽样, 共抽得 10 个单位 (其中有一个单位抽中 2 次). 对每个样本单位调查其所有货车完成的周转量与运量之和, 其数据列于表 6.7, 试

表 6.6 某部门各单位拥有的营业性货车的吨位和以及 PPS 抽样结果

单位 i	吨位和 M_i	累积吨位	单位 i	吨位和 M_i	累积吨位	单位 i	吨位和 M_i	累积吨位
1	104	104	17	115	5342	33	157	8147
2	88	192	18	148	5490	34	107	8254
3	26	218	19	164	5654	35	74	8328
4	542	760	20	60	5714	36	62	8390
5	117	877	21	246	5960	37	57	8447
6	55	932	22	162	6122	38	468	8915
7	2136	3068	23	30	6152	39	245	9160
8	179	3247	24	100	6252	40	136	9296
9	80	3327	25	124	6376	41	120	9416
10	740	4067	26	378	6754	42	216	9632
11	288	4355	27	89	6843	43	460	10092
12	168	4523	28	143	6986	44	955	11047
13	132	4655	29	145	7131	45	387	11434
14	200	4855	30	244	7375	46	64	11498
15	326	5180	31	527	7902	47	108	11606
16	47	5227	32	88	7990	48	255	11861

估计该部门全月完成的 $Y^{(1)}$ 与 $Y^{(2)}$, 并计算其精度.

10 个 1~11861 范围内的随机数(按产生的顺序)

及对应抽中的样本单位号如下:

5095(15), 10777(44), 7547(31), 2109(7), 9940(43),

6610(26), 9232(7), 4868(12), 8298(35), 467(4).

表 6.7 运输量调查的样本数据(按原单位序号顺序)

样本序号 i	原单位 序号	吨位和 $m_i(t)$	周 转 量 $y_i^{(1)}(t \cdot km)$	运量 $y_i^{(2)}(t)$	吨平均周 转量 $\bar{y}_i^{(1)}$	吨平均运 量 $\bar{y}_i^{(2)}$
1	4	542	2724040	14848	5025.904	27.39483
2	7	2136	9331140	48292	4368.511	22.60861
3	7	2136	9331140	48292	4368.511	22.60861
4	12	168	729790	4369	4343.988	26.00595
5	15	325	1547960	7485	4762.953	23.03076
6	26	378	1928600	8061	5102.116	21.92539
7	31	527	2182280	14436	4140.948	27.39278
8	35	74	319780	1756	4321.351	23.72973
9	43	460	2019930	10552	4391.152	22.93913
10	44	955	4754870	24635	4978.921	25.79581
				\bar{y}	4580.485	24.28316

根据表 6.7 中的样本数据, 按(6.33)与(6.36)式, 可计算总周转量与总运量的估计与方差, 进一步可计算估计量的标准差与变异系数.

$$M_0 = 11861, \quad n = 10,$$

$$\hat{Y}_{HH} = \frac{M_0}{n} \sum_{i=1}^n \frac{y_i}{m_i} = \frac{M_0}{n} \sum_{i=1}^n \bar{y}_i = M_0 \bar{y},$$

$$v(\hat{Y}_{HH}) = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2,$$

对于总周转量 $Y^{(1)}$ 的估计, 计算结果为:

$$\hat{Y}_{HH}^{(1)} = M_0 \bar{y}^{(1)} = 11861 \times 4580.485 = 54328549(t \cdot km),$$

$$v(\hat{Y}_{HH}^{(1)}) = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (y_i^{(1)} - \bar{y}^{(1)})^2 = \frac{11861^2}{10} \times 122725.5 \\ = 1.7265 \times 10^{12},$$

$$s(\hat{Y}_{HH}^{(1)}) = \sqrt{v(\hat{Y}_{HH}^{(1)})} = 1313980(t \cdot km),$$

$$cv(\hat{Y}_{HH}^{(1)}) = \frac{s(\hat{Y}_{HH}^{(1)})}{\hat{Y}_{HH}^{(1)}} = 2.4186\%.$$

对于总运量 $Y^{(2)}$ 的估计, 计算结果为:

$$\hat{Y}_{HH}^{(2)} = M_0 \bar{y}^{(2)} = 11861 \times 24.28316 = 288022.6(t),$$

$$\begin{aligned}
v(\hat{P}_{HH}^{(2)}) &= \frac{M_o^2}{n(n-1)} \sum_{i=1}^n (y_i^{(2)} - \bar{y}^{(2)})^2 = \frac{11861^2}{10} \times 4.738465 \\
&= 66662308, \\
s(\hat{P}_{HH}^{(2)}) &= \sqrt{v(\hat{P}_{HH}^{(2)})} = 8164.70(t), \\
cv(\hat{P}_{HH}^{(2)}) &= \frac{s(\hat{P}_{HH}^{(2)})}{\hat{P}_{HH}^{(2)}} = 2.8347\%.
\end{aligned}$$

从变异系数值可知, 对两个指标估计的相对误差在 95% 置信度下约为 5% (两倍变异系数)。

第7章

二阶与多阶抽样

§ 7.1 引言

7.1.1 定义及适用场合

定义 7.1 若总体中的 N 个初级单元每个都由若干次级(或称二级)单元组成, 在总体中按某种程序抽取 n 个初级单元, 然后对每个被抽中的初级单元再抽取若干个次级单元, 这种抽样称为二阶抽样, 也称二级抽样(two-stage sampling), 其中总体中抽取初级单元称为第一阶抽样, 从初级单元中抽取次级单元称为第二阶抽样。

从整群抽样中我们知道, 如果同一初级单元中的次级单元比较相似, 也即当群内相关 ρ_c 比较大时, 整群抽样的效率就比较低。事实上, 此时没有必要对该初级单元中的所有次级单元都进行调查, 仅需调查其中一部分即可。换言之, 此时需要在每个被抽中的初级单元中, 对次级单元进行一次再抽样, 这就是二阶抽样。

如果每个二级单元又可进一步分为更小的三级单元, 那么在每个第二阶抽样中被抽中的二级单元中, 若对其中的三级单元进行再抽样, 也即进行第三阶抽样, 则整个抽样过程就称为三阶抽样(three-stage sampling)。以此类推, 可以定义更一般的多阶抽样(multi-stage sampling)。

二阶及多阶抽样保持了(一阶)整群抽样样本单元相对集中的特点, 因此实施方便且平均每个基本单元的调查费用也较低。另一方面, 二阶与多阶抽样又避免了对较小单元进行过多调查的浪费, 因而大大提高了效率。多阶抽样的另一优点是在抽样时并不需要全部二级或更低级单元的抽样框。当然, 对于第一阶抽样, 初级单元的抽样框是必需的。在以后各阶抽样中, 仅仅需对那些已抽中的单元准备下一级单元的抽样框。这在实际问题中是非常方便的。因而多阶抽样(包括二阶抽样)在实际中应用非常广泛。特别是当抽样单元直接采用各级行政单位或有隶属关系的单位时, 更是如此。例如对于一项全国性抽样调查, 若调查不需要在每个省

进行时,就可将省作为一级单元,第一阶抽样先抽省,然后在每个抽中的省(或称样本省)进行第二阶抽样——抽市或县。对每个样本市、县又可进行第三阶抽样——抽街道、镇或乡等等。在这过程中,我们并不需要准备全国各省中的市、县及街道、乡、镇的抽样框。在第一阶抽样中,仅需要关于省(自治区、直辖市)的抽样框。对于每个被抽中的省(自治区、直辖市)才需要进一步准备市、县的抽样框,对每个被抽中的市、县准备有关街道及乡镇的抽样框,……。从行政系统而言,街道、乡、镇以下划分居民委员会或村民委员会,居(村)民委员会以下划分居(村)民小组,直到住户及住户中的每一口人。当然在多阶抽样中,各级单元的划分并不一定与行政系统完全一致,是比较灵活的。例如在全国性抽样中如果将市、县作为一级单元,则只要准备一份全国所有市县的抽样框,即可直接抽市、县。同样,对于一项以住户为基本单元的调查,可以在抽到街道或乡镇以后,跳过居(村)民委员会或居(村)民小组,直接抽户。再如对于一项在京中央直属单位的专业技术人员情况的调查,可将部委、司局级单位、处或基层单位与个人作为各级抽样单元,进行多阶抽样。显然,多阶抽样的组织管理也是比较方便的。

多阶抽样还可用于“散料”的抽样,即散料抽样(bulk sampling)。所谓“散料”,是指连续松散的、不易区分个体或抽样单元的材料。例如一堆煤,一仓库粮食,一列车水泥,一船化肥等。对于散料,抽样单元需要人为划分,当然也可以取其自然的单位,特别是当货物已经包装后。通常对于散料抽样,一级单元是自然或人为划分的分装(例如一袋化肥或一车皮矿石),二级单元则是从分装中(有时需要从其中各个部位)抽取一定数量(例如一公斤)的份样。

7.1.2 实施方法及同其他抽样方法的关系

在二阶或多阶抽样中,每一阶的具体抽样可以是多种多样的。在应用中,比较多的情形是当初级单元大小相等时,常用简单随机抽样;而当初级单元大小不等时,则在第一阶抽样时多用放回或不放回的与单元大小成比例的不等概率抽样。在每个阶段也可用下一章介绍的系统抽样。此外,二阶或多阶抽样也常与分层抽样及整群抽样结合起来。在许多情况是在某些阶的抽样(特别是第一阶与第二阶抽样)中进行分层抽样的。在某些情形,则是在最后一阶中不再抽样,调查下一级的所有单元,这即是多阶整群抽样(multi-stage cluster sampling)。例如:若在一个二阶抽

样中,对每个在第二阶抽样抽中的二级单元调查其中所有的三级单元,就是一个二阶整群抽样。

实际上,分层抽样与整群抽样都可以看成是多阶抽样的特例。以二阶抽样为例,若总体中包含 N 个初级单元,在第一阶抽样从中抽取 n 个初级单元,设第 i 个初级单元中包含 M_i 个次级单元,第二阶抽样从中抽取 m_i 个次级单元,则当 $n=N$ 时,即是分层抽样;而对每个 i ,当 $m_i=M_i$ 时,则是(一阶)整群抽样。

对于一个二阶或多阶抽样,抽样过程包括两步或多步,因此对于总体参数 θ 的任何一个估计量 $\hat{\theta}$ 求均值与方差时,必须用第3章中有关二步抽样的结果(引理3.1)。例如对于二阶抽样

$$E(\hat{\theta}) = E_1 E_2(\hat{\theta}), \quad (7.1)$$

$$V(\hat{\theta}) = V_1[E_2(\hat{\theta})] + E_1[V_2(\hat{\theta})], \quad (7.2)$$

其中 E_1, V_1 分别是对第一阶抽样求的均值与方差, E_2, V_2 是对固定的第一阶抽样中抽得的一组初级单元对第二阶抽样求的均值与方差。对于三阶抽样,也有类似的公式:

$$E(\hat{\theta}) = E_1 E_2 E_3(\hat{\theta}), \quad (7.3)$$

$$V(\hat{\theta}) = V_1\{E_2[E_3(\hat{\theta})]\} + E_1\{V_2[E_3(\hat{\theta})]\} + E_1\{E_2[V_3(\hat{\theta})]\}, \quad (7.4)$$

在本章中重点讨论二阶抽样。为了简化起见,先考虑每个初级单元都包含相等数量(M 个)的次级单元,在每个第一阶抽样中抽中的所有 n 个初级单元中抽取 m 个次级单元,且所用的抽样都是简单随机的。在此基础上,讨论一般的初级单元大小不等的情形。对此,我们先考虑 $n=1$ 的特殊情形,然后再推广到 $n>1$ 的一般情形。最后就三阶以及更高阶的多阶抽样作一简单的介绍。

§ 7.2 二阶抽样——初级单元大小相等的情形

本节讨论每个初级单元大小(即其中包含的次级单元的数目)相等情形的二阶抽样。我们假定每一阶抽样都是按简单随机抽样进行的。第一阶抽样是从 N 个初级单元中抽取 n 个初级单元;第二阶抽样是在每个抽中的初级单元所包含的 M 个次级单元中抽取 m 个次级单元。另外,第二阶抽样对每个初级单元而言都是相互独立的。

考虑这种情形的主要原因是为简化问题的讨论,得到二阶抽样的基

本结果。在实用中,可先将总体中的初级单元按大小分层,使层内的单元大小大致相同,从而可应用本节的结果。如果不能这样做或者为了使设计更为精确,则须用以后各节介绍的方法。

7.2.1 记号

记 Y_{ij} 为第 i 个初级单元中第 j 个次级单元的指标值 ($i=1, 2, \dots, N; j=1, 2, \dots, M$)。

y_{ij} 为样本中第 i 个初级单元中第 j 个次级单元的观测值 ($i=1, 2, \dots, n; j=1, 2, \dots, m$)。

$$f_1 = \frac{n}{N},$$

$$f_2 = \frac{m}{M},$$

$$Y_i = \sum_{j=1}^M Y_{ij},$$

$$y_i = \sum_{j=1}^m y_{ij},$$

$$\bar{Y}_i = Y_i / M,$$

$$\bar{y}_i = y_i / m,$$

$$\bar{Y} = \sum_{i=1}^N Y_i / N,$$

$$\bar{y} = \sum_{i=1}^n \bar{y}_i / n,$$

$$S_1^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2, \quad s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2,$$

$$S_2^2 = \frac{1}{N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2,$$

$$s_2^2 = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2.$$

其中 S_1^2 与 s_1^2 分别是总体与样本中初级单元间的方差, S_2^2 与 s_2^2 分别是总体与样本中同一初级单元中次级单元间的方差, 或称初级单元内的方差。如果令

$$S_{2i}^2 = \frac{1}{M-1} \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2,$$

则 S_2^2 即是所有 S_{2i}^2 的平均值:

$$S_2^2 = \frac{1}{N} \sum_{i=1}^N S_{2i}^2.$$

注意在第6章中, Y_i 与 y_i , \bar{Y}_i 与 \bar{y}_i 除了分别表示总体与样本的有关量外, 在数值上并无差别。而在本章中 y_i 仅仅是第二阶抽样中所抽得的这部分(共 m 个)次级单元观测值的和, \bar{y}_i 也只是这些单元观测值的平均值。

7.2.2 估计量及其方差

定理 7.1 对于二阶抽样, 若两个阶段的抽样都是简单随机的, 则

$$E(\bar{y}) = \bar{Y}, \quad (7.5)$$

$$V(\bar{y}) = \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{mn} S_2^2. \quad (7.6)$$

证明 由于每一阶抽样都是简单随机的, 根据(7.1)式, 有

$$\begin{aligned} E(\bar{y}) &= E_1 E_2 \left(\frac{1}{n} \sum_{i=1}^n \bar{y}_i \right) \\ &= E_1 \left[\frac{1}{n} \sum_{i=1}^n E_2(\bar{y}_i) \right] \\ &= E_1 \left[\frac{1}{n} \sum_{i=1}^n \bar{Y}_i \right] = \bar{Y}. \end{aligned}$$

根据(7.2)式, 注意到在每个初级单元中的第二阶抽样是相互独立的, 故有

$$\begin{aligned} V(\bar{y}) &= V_1[E_2(\bar{y})] + E_1[V_2(\bar{y})] \\ &= V_1 \left(\frac{1}{n} \sum_{i=1}^n \bar{Y}_i \right) + E_1 \left[V_2 \left(\frac{1}{n} \sum_{i=1}^n \bar{y}_i \right) \right] \\ &= \frac{1-f_1}{n} \cdot \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N-1} \\ &\quad + E_1 \left[\frac{1}{n^2} \sum_{i=1}^n \frac{1-f_2}{m} \cdot \frac{\sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2}{M-1} \right] \\ &= \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{mn} E_1 \left[\frac{1}{n} \sum_{i=1}^n S_{2i}^2 \right] \\ &= \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{mn} \left[\frac{1}{N} \sum_{i=1}^N S_{2i}^2 \right] \\ &= \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{mn} S_2^2. \quad \blacksquare \end{aligned}$$

定理 7.2 对于二阶抽样, 若两个阶段的抽样都是简单随机的, 则

$$v(\bar{y}) = \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{mn} s_2^2 \quad (7.7)$$

是 $V(\bar{y})$ 的无偏估计.

证明 对 s_1^2 与 s_2^2 , 分别利用(7.1)式分两步求均值:

$$E_2[(n-1)s_1^2] = E_2 \left[\sum_{i=1}^n (\bar{y}_i - \bar{y})^2 \right]$$

$$\begin{aligned}
&= \sum_{i=1}^n E_2(\bar{y}_i^2) - nE_2(\bar{y}^2) \\
&= \sum_{i=1}^n \{[E_2(\bar{y}_i)]^2 + V_2(\bar{y}_i)\} - n\{[E_2(\bar{y})]^2 + V_2(\bar{y})\} \\
&= \sum_{i=1}^n \bar{Y}_i^2 + \sum_{i=1}^n \frac{1-f_2}{m} S_{2i}^2 - n\left(\frac{1}{n} \sum_{i=1}^n \bar{Y}_i\right)^2 \\
&\quad - \frac{1-f_2}{nm} \sum_{i=1}^n S_{2i}^2 \\
&= \sum_{i=1}^n (\bar{Y}_i - \bar{\bar{Y}}_n)^2 + \frac{(n-1)(1-f_2)}{nm} \sum_{i=1}^n S_{2i}^2.
\end{aligned}$$

在上式中

$$\bar{\bar{Y}}_n \triangleq \frac{1}{n} \sum_{i=1}^n \bar{Y}_i \neq \bar{Y},$$

因而

$$\begin{aligned}
E(s_1^2) &= E_1[E_2(s_1^2)] \\
&= E_1\left[\frac{\sum_{i=1}^n (\bar{Y}_i - \bar{\bar{Y}}_n)^2}{n-1}\right] + \frac{1-f_2}{m} E_1\left[\frac{\sum_{i=1}^n S_{2i}^2}{n}\right] \\
&= S_1^2 + \frac{1-f_2}{m} S_2^2,
\end{aligned} \tag{7.8}$$

而

$$\begin{aligned}
E(s_2^2) &= E_1[E_2(s_2^2)] \\
&= E_1\left\{\frac{1}{n} \sum_{i=1}^n E_2\left[\frac{1}{m-1} \sum_{j=1}^m (y_{ij} - y_i)^2\right]\right\} \\
&= E_1\left\{\frac{1}{n} \sum_{i=1}^n \left[\frac{1}{M-1} \sum_{j=1}^M (\bar{Y}_{ij} - \bar{Y}_i)^2\right]\right\} \\
&= \frac{1}{N} \sum_{i=1}^M S_{1i}^2 = S_2^2,
\end{aligned} \tag{7.9}$$

从而

$$\begin{aligned}
E[v(\bar{y})] &= E\left[\frac{1-f_1}{n} s_1^2\right] + E\left[\frac{f_1(1-f_2)}{mn} s_2^2\right] \\
&= \frac{1-f_1}{n} S_1^2 + \frac{(1-f_1)(1-f_2) + f_1(1-f_2)}{mn} S_2^2 \\
&= \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{mn} S_2^2 = V(\bar{y}). \quad \blacksquare
\end{aligned}$$

从(7.8)式与(7.9)式可以看到对于此种情形的二阶抽样, 初级单元内样本方差 s_2^2 仍是总体相应方差 S_2^2 的无偏估计, 但样本初级单元间的方差 s_1^2 并不是总体初级单元间方差 S_1^2 的无偏估计. 因此 \bar{y} 方差估计公

式(7.7)在形式上与 $V(\bar{y})$ 的公式稍有差别, 在第二项中多一个 f_1 的因子. 这一点请读者特别注意.

由于(7.6)式与(7.7)式中第二项的系数比第一项要小得多, 因此在二阶抽样中, 估计量的方差的主项是第一项. 第二项与第一项比较起来, 通常要小很多. 参看 § 7.2.4 的数值例子.

推论 对于二阶抽样

$$\hat{S}_1^2 = s_1^2 + \frac{(1-f_2)s_2^2}{m} \quad (7.10)$$

是 S_1^2 的无偏估计.

证明 根据(7.8)与(7.9)式即得. ■

7.2.3 最优抽样比例

在二阶抽样中, 在给定总费用下如何确定第一阶抽样样本量 n 与第二阶抽样(每个初级单元中的抽样)的样本量 m , 使估计量 \bar{y} 的方差达到最小, 或在给定的 $V(\bar{y})$ 条件下, 使费用最省, 这就是最优抽样比 f_1 、 f_2 的确定问题.

考虑下述简单的线性费用函数

$$C = c_0 + c_1n + c_2nm, \quad (7.11)$$

若初级单元间的旅费不占重要位置, 则上述费用函数被证明是适用的. 这里 c_0 是与样本量无关的固定费用, c_1 、 c_2 分别是每调查一个初级单元与次级单元的费用. 注意(7.6)式可改写成

$$V(\bar{y}) = \frac{1}{n} \left(S_1^2 - \frac{S_2^2}{M} \right) + \frac{S_2^2}{mn} - \frac{S_1^2}{N}. \quad (7.12)$$

上式最后一项不依赖 m 与 n , 于是在固定 C 下极小化 $V(\bar{y})$ 或固定 $V(\bar{y})$ 下极小化 C 等价于极小化

$$\begin{aligned} \left(V + \frac{1}{N} S_1^2 \right) (C - c_0) &= \left[\left(S_1^2 - \frac{S_2^2}{M} \right) + \frac{S_2^2}{m} \right] (c_1 + c_2m) \\ &\triangleq \left(S_u^2 + \frac{S_2^2}{m} \right) (c_1 + c_2m), \end{aligned} \quad (7.13)$$

其中

$$S_u^2 = S_1^2 - \frac{S_2^2}{M}. \quad (7.14)$$

根据 Cauchy-Schwarz 不等式, 当下式

$$\frac{S_u}{S_2/\sqrt{m}} = \frac{\sqrt{c_1}}{\sqrt{c_2m}}$$

成立时, (7.13)式达到极小值. 因此 m 的最优值

$$m_{\text{opt}} = \frac{S_2}{S_u} \sqrt{\frac{c_1}{c_2}}. \quad (7.15)$$

由于 m_{opt} 一般不为整数, 在具体应用时, 应将它舍入成整数. 为此, Cameron(1951)给出了以下的规则:

令 m' 是 m_{opt} 的整数部分, 即 $m' = [m_{\text{opt}}]$, 则

(1) 若 $m_{\text{opt}}^2 > m'(m'+1)$, 则取 $m = m' + 1$;

(2) 若 $m_{\text{opt}}^2 \leq m'(m'+1)$, 则取 $m = m'$;

(3) 若 $m_{\text{opt}} > M$ 或 $S_1^2 - \frac{S_0^2}{M} < 0$, 则取 $m = M$.

求出 m 后, 根据(7.11)或(7.12)式即可求出 n 的值, 从而确定了最优的 f_1 与 f_2 .

可以证明, 当 n 大时,

$$\frac{S_2^2}{S_u^2} \approx \frac{1 - \rho_0}{\rho_0}, \quad (7.16)$$

其中 ρ_0 是将初级单元看作群的群内相关(系数).

7.2.4 数值例子——生猪存栏量的调查

例 7.1 为调查某县年终时生猪的存栏数量, 采用二阶抽样. 第一阶按简单随机抽样抽村, 第二阶在抽中的村中抽农户. 登记这些农户当时实养的生猪头数. 有关数据如下:

$$N = 325, \quad \bar{M} = 54 \text{ (平均每村的农户数)},$$

$$n = 12, \quad m = 10,$$

$$f_1 = 0.0369, \quad f_2 = 0.1852.$$

样本数据如表 7.1 所示.

$$\text{表 7.1 中的 } s_{2i}^2 = \frac{1}{m-1} \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2.$$

根据表 7.1 的数据, 可计算得:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i = 3.150,$$

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 = 2.323,$$

$$s_2^2 = \frac{1}{n} \sum_{i=1}^n s_{2i}^2 = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2 = 1.461,$$

从而

表 7.1 某县生猪存栏量调查的样本数据

i	y_{ij}	y_i	y_i	s_{2i}^2
1	3, 1, 2, 2, 1, 1, 1, 0, 2, 3	16	1.6	0.933
2	2, 1, 3, 0, 4, 2, 5, 2, 3, 2	24	2.4	2.044
3	7, 4, 6, 4, 6, 5, 6, 4, 3, 6	51	5.1	1.656
4	4, 5, 3, 2, 1, 4, 3, 2, 2, 5	31	3.1	1.878
5	5, 4, 5, 3, 7, 5, 4, 6, 5, 4	48	4.8	1.289
6	2, 1, 0, 2, 3, 1, 2, 0, 2, 1	14	1.4	0.933
7	2, 5, 4, 0, 2, 2, 2, 3, 3, 2	25	2.5	1.833
8	4, 4, 4, 3, 5, 2, 4, 6, 4, 5	41	4.1	1.211
9	1, 3, 2, 0, 2, 1, 2, 0, 3, 2	16	1.6	1.156
10	3, 5, 7, 4, 5, 3, 7, 4, 5, 4	47	4.7	2.011
11	0, 1, 1, 0, 3, 2, 1, 2, 1, 3	14	1.4	1.156
12	4, 5, 8, 5, 3, 4, 6, 5, 4, 5	51	5.1	1.433

$$\begin{aligned}
 v(\bar{y}) &= \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{mn} s_2^2 \\
 &= \frac{0.9631 \times 2.323}{12} + \frac{0.0369 \times 0.8148 \times 1.461}{12} \\
 &= 0.18644 + 0.00037 = 0.18681, \\
 \sqrt{v(\bar{y})} &= 0.4322.
 \end{aligned}$$

全县生猪存栏量的估计 \hat{Y} 及其标准差估计分别为:

$$\hat{Y} = 325 \times 54 \times 3.15 = 55283 \text{ (头)},$$

$$s(\hat{Y}) = \sqrt{v(\hat{Y})} = 325 \times 54 \times 0.4322 = 7588 \text{ (头)}.$$

从标准差的数值看, 估计量的精度是不够的. 因此需要加大样本量才能改善精度. 而从上面计算过程中可以看出, 需要加大的只是第一阶抽样的样本量 n , 也即抽村的数目. 因为方差 $v(\bar{y})$ 的主要来源是第一项. 事实上, 在本例中取 $m=10$ 未必合理. 设费用函数由 (7.11) 式给出, 即

$$C = c_0 + c_1 n + c_2 nm,$$

则按 (7.15) 式, m 的最优值由下式确定:

$$m_{\text{opt}} = \frac{S_2}{S_u} \sqrt{\frac{c_1}{c_2}}.$$

我们用 $\hat{S}_1^2 = s_1^2$ 与 $\hat{S}_2^2 = \frac{(1-f_2)s_2^2}{m}$ 分别估计 S_1^2 及 S_2^2 , 而

$$\hat{S}_1^2 = \hat{S}_1^2 - \frac{\hat{S}_2^2}{M} = s_1^2 - \frac{(1-f_2)s_2^2}{m} = \frac{s_1^2}{M} = 2.1769.$$

$$m_{\text{opt}} = \sqrt{\frac{1.461c_1}{2.177c_2}} = 0.8192\sqrt{\frac{c_1}{c_2}}.$$

设 $c_1/c_2 = 20$, 则 $m_{\text{opt}} = 3.66$.

因为 $m_{\text{opt}}^2 = 13.42 > 3 \times 4 = 12$, 故按 Cameron 规则取 $m = 4$, 固定 m 后, n 由总费用 C 或对 \bar{y} 的方差要求的数值而定.

7.2.5 关于比例的估计

若所有的次级单元可分成两类, 欲估计具有某种特性单元的比例, 则可用通常的方法, 令

$$Y_{ij} = \begin{cases} 1, & \text{若第 } i \text{ 初级单元及 } j \text{ 次级单元具有此特征;} \\ 0, & \text{否则.} \end{cases}$$

令

$$p_i = \frac{a_i}{m} \quad (7.17)$$

为抽到的第 i 个(初级)单元中具有此种特征的样本次级单元的比例, 则总体比例 P 的估计为:

$$\bar{p} = \sum_{i=1}^n p_i/n. \quad (7.18)$$

此时, 若记 $q_i = 1 - p_i$, 则

$$\begin{aligned} \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2 &= mp_iq_i, \\ s_1^2 &= \frac{1}{n-1} \sum_{i=1}^n (p_i - \bar{p})^2, \\ s_2^2 &= \frac{m}{n(m-1)} \sum_{i=1}^n p_iq_i. \end{aligned}$$

根据定理 7.1 与定理 7.2, 有以下定理:

定理 7.3 关于比例的二阶抽样, 若两个阶段的抽样都是简单随机的, 则

$$\bar{p} = \sum_{i=1}^n p_i/n$$

是总体比例 P 的无偏估计, 且 $V(\bar{p})$ 的一个无偏估计为:

$$v(\bar{p}) = \frac{1-f_1}{n(n-1)} \sum_{i=1}^n (p_i - \bar{p})^2 + \frac{f_1(1-f_2)}{n^2(m-1)} \sum_{i=1}^n p_iq_i. \quad (7.19)$$

例 7.2 对某市专业技术人员现状的调查中, 有这样一个问题: “您是否赞成单位有选择工作人员的权利, 同时工作人员有选择单位的权

利?”回答的选择项只有两个,即“赞成”(记为 1),“不赞成”(记为 0)。设抽样方案是按二阶抽样抽取的:第一阶抽单位,第二阶在抽中的单位中抽专业技术人员。为简单处理起见,暂且假定两阶抽样都是简单随机的,每个单位所包含的专业技术人员数目的差异也不大。有关参数如下:

$$n = 250, f_1 = 0.045, m = 5, f_2 = 0.12 \text{ (平均值)}.$$

在 250 个样本单位中,对该问题回答“赞成”的专业人员数 k ($k = 0, 1, \dots, 5$) 的频数 n_k 分布如表 7.2 所列。

表 7.2

k	0	1	2	3	4	5
n_k	1	8	38	57	125	21
p_k	0	0.2	0.4	0.6	0.8	1.0

于是

$$\begin{aligned}\bar{p} &= \sum n_k p_k \\ &= \frac{1}{250} [0 \times 1 + 0.2 \times 8 + 0.4 \times 38 + 0.6 \times 57 + 0.8 \times 125 + 1.0 \times 21] \\ &= \frac{172}{250} = 0.688,\end{aligned}$$

$$\begin{aligned}\sum_{k=0}^5 (p_k - \bar{p})^2 &= \sum_{k=0}^5 n_k p_k^2 - n \bar{p}^2 \\ &= 127.92 - 118.336 = 9.584,\end{aligned}$$

$$\sum_{k=0}^5 p_k q_k = \sum_{k=0}^5 n_k p_k q_k = 44.08.$$

于是

$$\begin{aligned}v(\bar{p}) &= \frac{1}{250} \frac{0.045}{(250 - 1)} \times 9.584 + \frac{0.045(1 - 0.12)}{250^2 \times 4} \times 44.08 \\ &= 0.000147 + 0.000007 = 0.000154, \\ \sqrt{v(\bar{p})} &= 0.0124.\end{aligned}$$

7.2.6 分层二阶抽样

对于分层二阶抽样,设同一层内的初级单元大小都相等,但不同层内的可以不相等,记第 h 层内每个初级单元包含 M_h 个次级单元,总体中的次级单元总数为 $\sum_{h=1}^L N_h M_h$ 个。在 h 层中按简单随机抽样抽 n_h 个初级单元,对每个被抽中的初级单元再用同样方式抽取 m_h 个次级单元,则总体

中按次级单元的均值 \bar{Y} 的分层二阶估计量为:

$$\bar{y}_{st} = \frac{\sum_h N_h M_h \bar{y}_h}{\sum_h N_h M_h} \triangleq \sum_h W_h \bar{y}_h, \quad (7.20)$$

其中

$$W_h = \frac{N_h M_h}{\sum_{h=1}^L N_h M_h} \quad (7.21)$$

是按次级单元个数的层权, 而

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{m_h} y_{hij}}{n_h m_h} \quad (7.22)$$

为 h 层的样本平均数,

将定理 7.1 与定理 7.2 用于每一层, 则有

$$V(\bar{y}_{st}) = \sum_h W_h^2 \left(\frac{1}{n_h} \frac{f_{1h}}{S_{1h}^2} + \frac{1-f_{2h}}{n_h m_h} \frac{S_{2h}^2}{S_{1h}^2} \right), \quad (7.23)$$

$$v(\bar{y}_{st}) = \sum_h W_h^2 \left(\frac{1}{n_h} \frac{f_{1h}}{S_{1h}^2} + \frac{f_{1h}(1-f_{2h})}{n_h m_h} \frac{S_{2h}^2}{S_{1h}^2} \right), \quad (7.24)$$

其中

$$f_{1h} = \frac{n_h}{N_h}, \quad f_{2h} = \frac{m_h}{M_h}. \quad (7.25)$$

为得到总体总量 Y 的估计 $\hat{Y}_{st} = \left(\sum_{h=1}^L N_h M_h \right) \bar{y}_h$ 的方差及其估计, 可在(7.23)式与(7.24)式中乘上 $(\sum_h N_h M_h)^2$.

与公式(7.15)一致, 在费用函数为

$$O = c_0 + \sum_h c_{1h} n_{1h} + \sum_h c_{2h} n_h m_h \quad (7.26)$$

时, 固定 O 使 V 达到极小或固定 V 使 O 达到极小的 m_h 的最优值为:

$$m'_h = \frac{S_{2h}}{\sqrt{S_{1h}^2 \cdot S_{2h}^2 / M_h}} \sqrt{\frac{c_{1h}}{c_{2h}}} \triangleq \frac{S_{2h}}{S_{1h}} \sqrt{\frac{c_{1h}}{c_{2h}}}. \quad (7.27)$$

现在我们考察在分层二阶抽样中, 所得样本是自加权的条件, 这个条件是在每层抽样中, 每个次级单元被抽中的概率皆相等, 或等价的, 对每一层总的抽样比 f_h 为常数 f_0 . 于是, 一个分层二阶样本是自加权的条件是:

$$f_{1h} f_{2h} = \frac{n_h}{N_h} \cdot \frac{m_h}{M_h} = f_0 \quad (h=1, 2, \dots, L). \quad (7.28)$$

根据(7.20)式, 对于自加权的分层二阶样本,

$$\bar{y}_{est} = \bar{y} \triangleq \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} y_{hij}}{\sum_{h=1}^L n_{hi} m_{hi}}. \quad (7.29)$$

§ 7.3 二阶抽样——初级单元大小不等情形($n=1$)

7.3.1 一般说明与记号

对大多数总体, 初级单元的大小不一定相等. 此时有两种处理方法: 第一种方法是将单元按大小分层, 使同一层中的单元大小相等或相差不多, 从而可用 7.2.6 段的方法处理. 但这种方法也有局限性, 可能分层后同一层的单元大小相差仍较大, 不能作为相等看待. 另一方面, 在一些实际问题中, 分层首先必须考虑其他的原则和因素, 从而不能照顾到单元的大小. 第二种处理方法正如对不等大小群的整群抽样那样, 对初级单元作不等概率抽样.

重新引进这一节及以后所用的记号如下:

Y_{ij} 表示第 i 个初级单元中第 j 个次级单元的观测值, 相应的样本值记为 y_{ij} ;

总体包含 N 个初级单元, 第一阶抽样的样本量为 n ;

对固定的初级单元, M_i 为其大小, 第二阶抽样的样本量为 m_{i1} ;

$$Y_i = \sum_{j=1}^{M_i} Y_{ij},$$

$$y_i = \sum_{j=1}^{m_{i1}} y_{ij},$$

$$\bar{Y}_i = Y_i / M_i,$$

$$\bar{y}_i = y_i / m_{i1},$$

$$S_{2i}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2,$$

$$s_{2i}^2 = \frac{1}{m_{i1} - 1} \sum_{j=1}^{m_{i1}} (y_{ij} - \bar{y}_i)^2.$$

对总体及所有二阶样本:

$$M_0 = \sum_{i=1}^N M_i, \quad m_0 = \sum_{i=1}^n m_{i1}, \quad \text{所包含的次级单元数;}$$

$$Y = \sum_{i=1}^N Y_i, \quad y = \sum_{i=1}^n y_i, \quad \text{总和;}$$

$$\bar{Y} = Y / M_0, \quad \bar{y} = y / \sum_{i=1}^n m_{i1}, \quad \text{按次级单元平均;}$$

$$\bar{Y} = Y / N, \quad \bar{y} = y / n, \quad \text{按初级单元平均;}$$

在本节中, 我们首先考虑 $n=1$ 的特殊情形, 即在总体中只抽取一个初级单元. 假定被抽中的初级单元为 i , 第二阶抽样是从其中 M_i 个次级

单元中按简单随机抽样抽取 m_i 个次级单元. 为方便起见, 我们主要考虑对 \bar{Y} 的估计. 为简便起见, 有时我们将初级单元简称为单元.

7.3.2 等概率抽取初级单元

设唯一样本单元 i 是根据等概率原则随机抽取的, 则可以考虑以下两种估计方法:

一、方法 I

估计量取为样本平均数:

$$\hat{y}_I = \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}, \quad (7.30)$$

$$E(\bar{y}_I) = E_1 E_2(\bar{y}_i) = E_1(\bar{Y}_i) = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i \triangleq \bar{Y}_o. \quad (7.31)$$

\bar{Y}_o 是 \bar{Y}_i 的不加权平均, 不等于 $\bar{Y} = \sum_{i=1}^N M_i \bar{Y}_i / M_0$, 因此 \hat{y}_I 是有偏的.

$$\begin{aligned} V(\hat{y}_I) &= V_1[E_2(\bar{y}_i)] + E_1[V_2(\bar{y}_i)] \\ &= V_1(\bar{Y}_i) + E_1\left[\frac{M_i - m_i}{M_i} \cdot \frac{S_{2i}^2}{m_i}\right] \\ &= \frac{1}{N} \sum_{i=1}^N (\bar{Y}_i - \bar{Y}_o)^2 + \frac{1}{N} \sum_{i=1}^N \frac{(M_i - m_i) S_{2i}^2}{M_i m_i}, \end{aligned} \quad (7.32)$$

于是

$$\text{MSE}(\bar{Y}_I) = (\bar{Y} - \bar{Y}_o)^2 + \frac{1}{N} \sum_{i=1}^N (\bar{Y}_i - \bar{Y}_o)^2 + \frac{1}{N} \sum_{i=1}^N \frac{(M_i - m_i) S_{2i}^2}{M_i m_i}. \quad (7.33)$$

上式中的第一项为偏倚的平方, 第二项为初级单元平均数 \bar{Y}_i 之间的差异, 第三项为初级单元内次级单元间的差异. 注意到根据(7.33)式, m_i 的选择对 $\text{MSE}(\hat{y}_I)$ 的前两项无关. 一般采用以下两种方法: 一是取 $m_i = m$ 为常数; 二是取 m_i 与 M_i 成比例.

二、方法 II

估计量取为:

$$\hat{y}_{II} = \frac{M_i}{M} \bar{y}_i = \frac{N M_i}{M_0} \bar{y}_i, \quad (7.34)$$

其中 $\bar{M} = M_0 / N$ 是(初级)单元的平均大小.

$$E(\hat{y}_{II}) = E_1\left[\frac{M_i}{M} \bar{Y}_i\right] = \frac{1}{N \bar{M}} \sum_{i=1}^N M_i \bar{Y}_i = \frac{1}{M_0} \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij} = \bar{Y}, \quad (7.35)$$

从而 \bar{y}_{II} 是无偏的.

$$\begin{aligned}
 V(\bar{y}_{II}) &= V_1[E_2(\bar{y}_{II})] + E_1[V_2(\bar{y}_{II})] \\
 &= V_1\left[\frac{M_i}{\bar{M}} \bar{Y}_i\right] + E_1\left[\frac{M_i^2}{\bar{M}^2} \frac{(M_i - m_i)S_{2i}^2}{m_i}\right] \\
 &= \frac{1}{N} \sum_{i=1}^N \left(\frac{M_i}{\bar{M}} \bar{Y}_i - \bar{Y}\right)^2 + \frac{1}{NM^2} \sum_{i=1}^N \frac{M_i(M_i - m_i)S_{2i}^2}{m_i} \\
 &= \frac{1}{N} \sum_{i=1}^N \left(\frac{Y_i}{\bar{M}} - \frac{Y}{M_0}\right)^2 + \frac{1}{NM^2} \sum_{i=1}^N \frac{M_i(M_i - m_i)S_{2i}^2}{m_i} \\
 &= \frac{1}{NM_0^2} \sum_{i=1}^N (NY_i - Y)^2 + \frac{N}{M_0^2} \sum_{i=1}^N \frac{M_i(M_i - m_i)S_{2i}^2}{m_i} \\
 &= \frac{N}{M_0^2} \sum_{i=1}^N (Y_i - \bar{Y})^2 + \frac{N}{M_0^2} \sum_{i=1}^N \frac{M_i(M_i - m_i)S_{2i}^2}{m_i}. \quad (7.36)
 \end{aligned}$$

上式表明, 当以 y_{II} 为估计量时, 初级单元的差异对方差的作用以单元总和 Y_i (与其平均数 \bar{Y}) 差异的形式出现. 如果 M_i 相差较大, 而 \bar{Y}_i 相对比较稳定时, Y_i 的差异就较大. 而这正是大多数实际总体的情形, 因而 $V(\bar{y}_{II})$ 常比 $MSE(\bar{y}_I)$ 还大. 因此虽然 \bar{y}_{II} 是无偏的, 但效果一般并不好.

7.3.3 不等概率抽取初级单元

设唯一样本单元 ϕ 是按一定概率从总体中抽取的, 我们考虑以下三种方法:

一、方法 III

抽样是按照单元大小 M_i 成正比的概率, 也即按 M_i/M_0 的概率抽取的 (PPS 抽样), 估计量为:

$$\bar{y}_{III} = \bar{y}_i, \quad (7.37)$$

$$E(\bar{y}_{III}) = E_1(\bar{Y}_i) = \sum_{i=1}^N \frac{M_i}{M_0} \bar{Y}_i = \bar{Y}, \quad (7.38)$$

$$\begin{aligned}
 V(\bar{y}_{III}) &= V_1(\bar{Y}_i) + E_1\left[\frac{(M_i - m_i)S_{2i}^2}{M_i m_i}\right] \\
 &= \sum_{i=1}^N \frac{M_i}{M_0} (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^N \frac{(M_i - m_i)S_{2i}^2}{M_0 m_i} \\
 &= \frac{1}{M_0} \left[\sum_{i=1}^N M_i (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^N \frac{(M_i - m_i)S_{2i}^2}{m_i} \right]. \quad (7.39)
 \end{aligned}$$

这表明 \bar{y}_{III} 是无偏的, 而当每个单元按次级单元平均差别不大时, 方

差一般也不会很大。

二、方法 IV

抽样是按指定的一组概率 Z_i 抽取的, $\sum_{i=1}^N Z_i = 1$, 估计量为:

$$\bar{y}_{IV} = \frac{M_i}{M_0} \frac{\bar{y}_i}{z_i}, \quad (7.40)$$

$$E(\bar{y}_{IV}) = E_1 \left[\frac{M_i}{M_0} \frac{Y_i}{z_i} \right] = \sum_{i=1}^N \frac{M_i}{M_0} Y_i = \bar{Y}, \quad (7.41)$$

$$\begin{aligned} V(\bar{y}_{IV}) &= V_1[E_2(\bar{y}_{IV})] + E_1[V_2(\bar{y}_{IV})] \\ &= V_1 \left(\frac{M_i}{M_0} \frac{\bar{Y}_i}{z_i} \right) + E_1 \left[\frac{M_i^2 (M_i - m_i) S_{2i}^2}{M_0^2 z_i^2 M_i m_i} \right] \\ &= \sum_{i=1}^N Z_i \left[\frac{M_i}{M_0} \frac{\bar{Y}_i}{z_i} - \bar{Y} \right]^2 + \sum_{i=1}^N \frac{M_i (M_i - m_i) S_{2i}^2}{M_0^2 Z_i m_i} \\ &= \frac{1}{M_0^2} \left[\sum_{i=1}^N Z_i \left(\frac{M_i Y_i}{Z_i} - M_0 \bar{Y} \right)^2 + \sum_{i=1}^N \frac{M_i (M_i - m_i) S_{2i}^2}{Z_i m_i} \right]. \end{aligned} \quad (7.42)$$

在此情况, 若取 $Z_i = M_i / M_0$, 则 $\bar{y}_{IV} = \bar{y}_{III}$; 若取 $Z_i = 1/N$, 则 $\bar{y}_{IV} = \bar{y}_{II}$.

三、方法 V

抽样是按指定一组概率 Z_i 抽取的, $\sum_{i=1}^N Z_i = 1$, 估计量为:

$$\bar{y}_V = y_i, \quad (7.43)$$

$$E(\bar{y}_V) = \sum_{i=1}^N Z_i \bar{Y}_i \triangleq \bar{Y}_z. \quad (7.44)$$

一般的, $\bar{Y}_z \neq \bar{Y}$, 故 \bar{y}_V 是有偏的。但当 $Z_i \approx M_i / M_0$ 时, 偏倚很小。与 (7.33) 式类似, \bar{y}_V 的均方误差可表示为:

$$\text{MSE}(\bar{y}_V) = (\bar{Y}_z - \bar{Y})^2 + \sum_{i=1}^N Z_i (\bar{Y}_i - \bar{Y}_z)^2 + \sum_{i=1}^N \frac{Z_i (M_i - m_i) S_{2i}^2}{M_i m_i}. \quad (7.45)$$

上式中三项的意义也与 (7.33) 式相应的项的意义相同。

例 7.3 为对以上 5 种方法进行比较, 考虑对以下一个人为总体 ($N=3$) 进行抽样 (此例引自 Cochran (1977))。

$$\bar{Y} = \frac{Y}{M_0} = \frac{33}{12} = 2.75,$$

$$\bar{Y}_z = \frac{1}{3} (0.5 + 2.0 + 4.0) = 2.167.$$

按方法 I ~ 方法 V 的抽样与估计方法抽一个单元的主要结果列于表

表 7.3 $N=3$ 初级单元不等大小的一个人为总体

初级单元:	Y_{ij}	M_i	Y_i	S_{2i}^2	\bar{Y}_i
1	0, 1	2	1	0.500	0.5
2	1, 2, 2, 3	4	8	0.667	2.0
3	3, 3, 4, 4, 5, 5	6	24	0.800	4.0
		$M_0=12$	$Y=33$		

7.4 中, 其中方法 I 又分 i) $m_i=2$ 与 ii) $m_i=M_i/2$ 两种情况, 其余几种方法, m_i 皆取为 2. 方法 IV 及 V 中的 Z_i 取为 (0.2, 0.4, 0.4) 是 M_i/M_0 ($i=1, 2, 3$) 的估计.

表 7.4 $n=1$ 时五种抽样和估计方法的比较

方法	抽取单元的概率 (括号中的数字为 例中的实际数字)	\bar{Y} 的估 计量	无偏 性	根据表 7.3 总体抽样的 MSE			
				(偏倚) ²	单元之间	单元内	总计
I ₀	$\frac{1}{N} \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right)$	$\left\{ \begin{array}{l} \bar{y}_i \\ \frac{NM_i}{M_0} \bar{y}_i \end{array} \right.$	有偏	0.340	2.056	0.144	2.541
I ₀₀	$\frac{1}{N} \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right)$			0.340	2.056	0.189	2.579
II	$\frac{1}{N} \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right)$	$\frac{NM_i}{M_0} \bar{y}_i$	无偏	0	5.792	0.256	6.048
III	$\frac{M_i}{M_0} (0.17, 0.33, 0.50)$	\bar{y}_i	无偏	0	1.813	0.189	2.002
IV	$Z_i (0.2, 0.4, 0.4)$	$\frac{M_i}{M_0} \frac{\bar{y}_i}{Z_i}$	无偏	0	3.583	0.213	3.796
V	$Z_i (0.2, 0.4, 0.4)$	\bar{y}_i	有偏	0.062	1.800	0.173	2.035

从表 7.4 的最后一列 MSE 的值可看出, 方法 II 的效果最差(虽然它无偏), 方法 III 最好. 方法 IV 与 V 的效果取决于 Z_i 的选择与 M_i/M_0 的符合程度. 方法 IV 虽然无偏, 但 MSE 并不太小. 上述结论虽然是对一个具体人造总体得出的, 但具有普遍意义.

§ 7.4 二阶抽样——初级单元大小不等的一般情形($n > 1$)

初级单元大小不等的一般情形既是上节 $n=1$ 情形的推广, 也是第 6 章中群大小不等情形的整群抽样 (§ 6.4) 的发展. 一个自然而基本的假定是第二阶抽样对不同的初级单元是相互独立的. 我们暂且假定第二阶

抽样都是简单随机的, 即对第一阶抽样中被抽中的第 i 个初级单元, 用不放回等概率抽样抽取 m_i 个次级单元, 并令

$$f_{2i} = \frac{m_i}{M_i} \quad (i = 1, 2, \dots, n).$$

不过后面这个假定可用任意一种其他抽样代替, 而结果没有实质性的变化(除非第二阶抽样是整群抽样, 这样整个抽样是单阶整群抽样而不是一般意义的二阶抽样). 因此, 在这一节中仍将重点放在初级单元的抽取方法. 为表示便利起见, 我们改以总体总和 Y 为估计的目标量.

7.4.1 按多项抽样抽取初级单元

设初级单元是按多项抽样抽取的, 即以给定的一组概率 Z_i ($\sum_{i=1}^N Z_i = 1$) 逐个放回独立抽取的, 重复 n 次, 共抽得 n 个(可能有重复)单元. 若有单元被重复抽中一次以上, 则原来在第二阶抽样中被抽中的 m_i 个次级单元也被放回, 按简单随机抽样重抽 m_i 个次级单元.

仿照第6章中的处理方法, 我们首先对 Y_i 作出估计: $\hat{Y}_i = M_i \bar{y}_i$, 然后用 Hansen-Hurwitz 估计量对 Y 作出估计, 形式如下:

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{Y}_i}{Z_i} = \frac{1}{n} \sum_{i=1}^n \frac{M_i \bar{y}_i}{Z_i}. \quad (7.46)$$

与定理 5.1 和 5.2 所用的方法类似, 由于在此情形, 第一阶抽样可看作是从“总体” $\{\hat{Y}_i/Z_i \mid (i = 1, 2, \dots, N)\}$ 中独立抽取的样本量为 n 的样本. 而 \hat{Y}_{HH} 是样本平均数, 因而 \hat{Y}_{HH} 的均值等于该“总体”的均值 Y , 从而它是无偏的. 它的方差为“总体”方差的 $1/n$, 即

$$V(\hat{Y}_{HH}) = \frac{1}{n} V\left(\frac{\hat{Y}_i}{Z_i}\right),$$

其中 $V\left(\frac{\hat{Y}_i}{Z_i}\right)$ 也即“总体”方差即是 $n=1$ 时用估计量 \hat{Y}_i/Z_i 估计 Y 的方差, 后者即等于上节中的方法 IV. 不过因为这里讨论的目标量是 Y 而不是 \bar{Y} 的估计, 因此

$$V\left(\frac{\hat{Y}_i}{Z_i}\right) = V\left(\frac{M_i \bar{y}_i}{Z_i}\right) = V(M_i \bar{y}_{IV}) = M_i^2 V(\bar{y}_{IV}).$$

于是根据(7.42)式有

$$V(\hat{Y}_{HH}) = \frac{1}{n} \left[\sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 + \sum_{i=1}^N \frac{M_i^2 (1 - f_{2i}) S_{2i}^2}{m_i Z_i} \right]. \quad (7.47)$$

如果不指定第二阶抽样方法, 则上式可改写成更一般的形式:

$$V(\hat{P}_{HH}) = \frac{1}{n} \left[\sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 + \sum_{i=1}^N \frac{V_2(\hat{P}_i)}{Z_i} \right], \quad (7.48)$$

而 $V(\hat{P}_{HH})$ 的估计可直接用 $\frac{\hat{P}_i}{Z_i}$ 样本方差的 $\frac{1}{n}$, 即

$$v(\hat{P}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{\hat{P}_i}{z_i} - \hat{P}_{HH} \right)^2, \quad (7.49)$$

它是 $V(\hat{P}_{HH})$ 的一个无偏估计, 从而得到如下定理:

定理 7.3 对于二阶抽样, 若第一阶抽样按放回的多项抽样抽取初级单元, 每次第 i 个单元入样概率为 $Z_i \left(\sum_{i=1}^N Z_i = 1 \right)$, \hat{P}_i 是第二阶抽样中对第 i 个初级单元总和 Y_i 的无偏估计, $V_2(\hat{P}_i)$ 是其方差, 则总体总和 Y 的一个无偏估计为:

$$\hat{P}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{\hat{P}_i}{z_i} = \frac{1}{n} \sum_{i=1}^n \frac{M_i \bar{y}_i}{z_i}.$$

它的方差为 $V(\hat{P}_{HH}) = \frac{1}{n} \left[\sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 + \sum_{i=1}^N \frac{V_2(\hat{P}_i)}{Z_i} \right]$,

而 $v(\hat{P}_{HH}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{\hat{P}_i}{z_i} - \hat{P}_{HH} \right)^2$

是 $V(\hat{P}_{HH})$ 的无偏估计.

本定理也可直接按二阶抽样求均值的一般公式用代数方法证明, 但推导过程稍为复杂些, 我们仅对 $v(\hat{P}_{HH})$ 的无偏性加以证明如下:

事实上, 注意到 \hat{P}_{HH} 的表达式以及 $E_2(\hat{P}_i) = Y_i$, 则有

$$\begin{aligned} E[v(\hat{P}_{HH})] &= \frac{1}{n(n-1)} \sum_{i=1}^n E \left[\left(\frac{Y_i}{z_i} - Y \right) - \left(\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{z_i} - Y \right) \right. \\ &\quad \left. + \left(\frac{\hat{P}_i}{z_i} - \frac{Y_i}{z_i} \right) - \left(\hat{P}_{HH} - \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{z_i} \right) \right]^2 \\ &= \frac{1}{n(n-1)} \sum_{i=1}^n \left[E \left(\frac{Y_i}{z_i} - Y \right)^2 + E \left(\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{z_i} - Y \right)^2 \right. \\ &\quad \left. + E \left(\frac{\hat{P}_i}{z_i} - \frac{Y_i}{z_i} \right)^2 + E \left(\hat{P}_{HH} - \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{z_i} \right)^2 \right. \\ &\quad \left. - 2E \left(\frac{Y_i}{z_i} - Y \right) \left(\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{z_i} - Y \right) \right. \\ &\quad \left. + 2E \left(\frac{Y_i}{z_i} - Y \right) \left(\frac{\hat{P}_i}{z_i} - \frac{Y_i}{z_i} \right) \right. \\ &\quad \left. - 2E \left(\frac{Y_i}{z_i} - Y \right) \left(\hat{P}_{HH} - \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{z_i} \right) \right] \end{aligned}$$

$$\begin{aligned}
& -2E\left(\frac{1}{n}\sum_{i=1}^n\frac{Y_i}{z_i}-Y\right)\left(\frac{\hat{P}_i}{z_i}-\frac{Y_i}{z_i}\right) \\
& +2E\left(\frac{1}{n}\sum_{i=1}^n\frac{Y_i}{z_i}-Y\right)\left(\hat{P}_{HH}-\frac{1}{n}\sum_{i=1}^n\frac{Y_i}{z_i}\right) \\
& -2E\left(\frac{\hat{P}_i}{z_i}-\frac{Y_i}{z_i}\right)\left(\hat{P}_{HH}-\frac{1}{n}\sum_{i=1}^n\frac{Y_i}{z_i}\right) \\
& =\frac{1}{n-1}\left\{E_1\left[\frac{1}{n}\sum_{i=1}^n\left(\frac{Y_i}{z_i}-Y\right)^2\right]\right. \\
& +E_1\left[\frac{1}{n}\sum_{i=1}^n\frac{Y_i}{z_i}-Y\right]^2+E_1E_2\left[\frac{1}{n}\sum_{i=1}^n\left(\frac{\hat{P}_i}{z_i}-\frac{Y_i}{z_i}\right)^2\right] \\
& +E_1E_2\left(\hat{P}_{HH}-\frac{1}{n}\sum_{i=1}^n\frac{Y_i}{z_i}\right)^2 \\
& -2E_1\left(\frac{1}{n}\sum_{i=1}^n\frac{Y_i}{z_i}-Y\right)^2 \\
& \left.-2E_1E_2\left(\hat{P}_{HH}-\frac{1}{n}\sum_{i=1}^n\frac{Y_i}{z_i}\right)^2\right\} \\
& =\frac{1}{n-1}\left\{E_1\left[\frac{1}{n}\sum_{i=1}^n\left(\frac{Y_i}{z_i}-Y\right)^2\right]\right. \\
& -V_1\left(\frac{1}{n}\sum_{i=1}^n\frac{Y_i}{z_i}\right)+E_1\left[\frac{1}{n}\sum_{i=1}^nV_2\left(\frac{\hat{P}_i}{z_i}\right)\right] \\
& \left.-E_1\left[\frac{1}{n^2}\sum_{i=1}^nE_2\left(\frac{\hat{P}_i}{z_i}-\frac{Y_i}{z_i}\right)^2\right]\right\} \\
& =\frac{1}{n-1}\left[\sum_{i=1}^NZ_i\left(\frac{Y_i}{Z_i}-Y\right)^2-\frac{1}{n}\sum_{i=1}^NZ_i\left(\frac{Y_i}{Z_i}-Y\right)^2\right. \\
& \left.+\sum_{i=1}^NV_2\left(\frac{\hat{P}_i}{Z_i}\right)-\frac{1}{n}\sum_{i=1}^NV_2\left(\frac{\hat{P}_i}{Z_i}\right)\right] \\
& =\frac{1}{n}\sum_{i=1}^NZ_i\left(\frac{Y_i}{Z_i}-Y\right)^2+\frac{1}{n}\sum_{i=1}^NV_2\left(\frac{\hat{P}_i}{Z_i}\right)=V(\hat{P}_{HH}). \blacksquare
\end{aligned}$$

如果需要对 $V(\hat{P}_{HH})$ 的两个分量分别进行估计, 则显然对第二阶抽样方差分量 $\frac{1}{n}\sum_{i=1}^NV_2\left(\frac{\hat{P}_i}{Z_i}\right)$ 的一个无偏估计是:

$$v_2(\hat{P}_{HH})=\frac{1}{n^2}\sum_{i=1}^n\frac{v_2(\hat{P}_i)}{z_i^2}, \quad (7.50)$$

其中 $v_2(\hat{P}_i)$ 是 $V_2(\hat{P}_i)$ 的一个无偏估计. 于是对第一阶抽样方差分量 $\frac{1}{n}\sum_{i=1}^NZ_i\left(\frac{Y_i}{Z_i}-Y\right)^2$ 的一个无偏估计为:

$$v_1(\hat{P}_{HH})=\frac{1}{n(n-1)}\sum_{i=1}^n\left(\frac{\hat{P}_i}{z_i}-\hat{P}_{HH}\right)^2-\frac{1}{n^2}\sum_{i=1}^n\frac{v_2(\hat{P}_i)}{z_i^2}. \quad (7.51)$$

下面我们求 \hat{P}_{HH} 是自加权的条件, 注意到 \hat{P}_{HH} 可表成以下形式:

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{M_i}{z_i m_i} \sum_{j=1}^{m_i} y_{ij}, \quad (7.52)$$

因此只有当 $M_i/(nz_i m_i)$ 皆相等, 为一常数时, 即

$$\frac{M_i}{nz_i m_i} = K = \frac{1}{f_0} \quad (7.53,$$

时, 有

$$\hat{Y}_{HH} = K \sum_{i=1}^n \sum_{j=1}^{m_i} y_{ij}.$$

此时估计量是自加权的, 这时 f_0 是总体中任意一个次级单元被抽中的概率, 也即总的抽样比. 在实际应用中, 若 f_0 事先确定, 则

$$f_0 = \frac{m_i}{M_i} = \frac{f_0}{nz_i} \quad (7.54)$$

可按已被抽中的初级单元确定.

对自加权样本, 估计量的方差估计也有以下简单的形式:

$$v(\hat{Y}_{HH}) = \frac{n}{(n-1)f_0^2} \sum_{i=1}^n (y_i - \bar{y})^2, \quad (7.55)$$

其中

$$y_i = \sum_{j=1}^{m_i} y_{ij}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

对每个 i , 若令 $Z_i = \frac{M_i}{M_0}$, 即若对初级单元进行 PPS 抽样, 则估计量可简化为:

$$\hat{Y}_{PPS} = \frac{M_0}{n} \sum_{i=1}^n \bar{y}_i. \quad (7.56)$$

若进而 $m_i = m$, 则样本是自加权的, 此时

$$\hat{Y}_{PPS} = M_0 \bar{y}, \quad (7.57)$$

其中

$$\bar{y} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m y_{ij}. \quad (7.58)$$

它也是 \bar{Y} 的无偏估计, $V(\hat{Y}_{PPS})$ 的一个无偏估计为:

$$v(\hat{Y}_{PPS}) = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2. \quad (7.59)$$

在这一节中我们考虑的第一阶抽样是放回抽样. 为保持其独立性, 前面规定当一个初级单元被重复抽中时, 前一次在第二阶抽样中被抽到的 m_i 个次级单元应放回重抽. 在实际执行中, 这样做是不方便的, 也没有此必要. 有以下两种变通方法可供选用. 设第 i 个初级单元被抽中 t_i 次, 第一种方法是用简单随机抽样在该单元中一次抽取 $m_i t_i$ (假定 $m_i t_i \leq M_i$) 个次级单元; 第二种方法是只在该单元中一次抽 m_i 个次级单元. 两种方

法的估计量皆取以下形式:

$$\sum_{i=1}^N t_i M_i \bar{y}_i / n Z_i.$$

此时,按第一种方法,实际方差比标准的(7.4式)减小,可以证明减少的量为 $\frac{n-1}{n} \sum_{i=1}^N M_i S_{2i}^2$; 而按第二种方法,实际方差比标准的要大,增加量为 $\frac{n-1}{n} \sum_{i=1}^N \frac{M_i^2 (1-f_{2i}) S_{2i}^2}{m_i}$.

7.4.2 不放回抽样时的一般结果

本段讨论当初级单元是按照某种方式不放回抽样时,二阶抽样的估计量、估计量的方差及其估计的一般结果,主要给出两个定理(定理7.4与7.5),这两个定理是由 Durbin (1953) 首先提出的,后经 Des Raj (1966), J. N. K. Rao (1975) 等推广发展的. 这里叙述的结果与证明是根据 Des Raj 的形式,而定理7.5的推论则是根据 Rao 的结果. 根据这两个定理,可以方便地将单阶抽样的结果移植到二阶甚至多阶抽样中.

我们考虑的基本假定是:初级单元是按照某种方式从总体中不放回地抽取的. 对单元 i , \hat{P}_i 是 Y_i 的无偏估计, $\hat{\sigma}_{2i}^2 = v_2(\hat{P}_i)$ 是 $\sigma_{2i}^2 = V_2(\hat{P}_i)$ 的无偏估计. 此外,对不同的 i , 第二阶抽样是相互独立的.

考虑 Y 的以下线性估计:

$$\hat{P} = \sum_{i=1}^n w_{is} \hat{P}_i. \quad (7.60)$$

这里的 w_{is} 既依赖于被抽中的单元 i , 也可能依赖于样本中的其他单元. 引进随机变量:

$$w'_{is} = \begin{cases} w_{is}, & \text{若单元 } i \text{ 入样;} \\ 0, & \text{否则.} \end{cases} \quad (7.61)$$

则

$$\hat{P} = \sum_{i=1}^N w'_{is} \hat{P}_i. \quad (7.62)$$

定理 7.4 在上述假定和记号下,

$$\hat{P} = \sum_{i=1}^N w'_{is} \hat{P}_i$$

是无偏的充要条件为:

$$E_1(w'_{is}) = 1, \text{ 对所有的 } i. \quad (7.63)$$

此时

$$V(\hat{P}) = V\left[\sum_{i=1}^n w_{is} \hat{P}_i\right] = V\left(\sum_{i=1}^n w_{is} Y_i\right) + \sum_{i=1}^n E_1(w'_{is}) \sigma_{2i}^2. \quad (7.64)$$

证明

$$E(\hat{P}) = E_1 E_2\left(\sum_{i=1}^N w'_{is} \hat{P}_i\right) = E_1\left(\sum_{i=1}^N w'_{is} Y_i\right) = \sum_{i=1}^N E_1(w'_{is}) Y_i.$$

因此, $E(\hat{P}) = \sum_{i=1}^N Y_i = Y$ 的充要条件是: 对所有的 i , 都有 $E_1(w'_{is}) = 1$.

$$\begin{aligned} V(\hat{P}) &= V_1[E_2(\hat{P})] + E_1[V_2(\hat{P})] \\ &= V_1\left(\sum_{i=1}^n w_{is} Y_i\right) + E_1\left[\sum_{i=1}^N w'_{is} V_2(\hat{P}_i)\right] \\ &= V\left(\sum_{i=1}^n w_{is} Y_i\right) + \sum_{i=1}^N E_1(w'_{is}) \sigma_{2i}^2. \quad \blacksquare \end{aligned}$$

定理 7.5 若二次型

$$v\left(\sum_{i=1}^n w_{is} Y_i\right) = \sum_{i=1}^n a_{is} Y_i^2 + 2 \sum_{i=1}^n \sum_{j=1}^n b_{ijs} Y_i Y_j \quad (7.65)$$

是 $V\left(\sum_{i=1}^n w_{is} Y_i\right)$ 的一个无偏估计, 则

$$v\left(\sum_{i=1}^n w_{is} \hat{P}_i\right) = \sum_{i=1}^n a_{is} \hat{P}_i^2 + 2 \sum_{i=1}^n \sum_{j=1}^n b_{ijs} \hat{P}_i \hat{P}_j + \sum_{i=1}^n w_{is} \hat{\sigma}_{2i}^2 \quad (7.66)$$

是 $V(\hat{P}) = V\left(\sum_{i=1}^n w_{is} \hat{P}_i\right)$ 的无偏估计.

证明 令

$$\begin{aligned} a'_{is} &= \begin{cases} a_{is}, & \text{若单元 } i \text{ 入样;} \\ 0, & \text{否则.} \end{cases} \\ b'_{ijs} &= \begin{cases} b_{ijs}, & \text{若单元 } i, j \text{ 都入样;} \\ 0, & \text{否则.} \end{cases} \end{aligned}$$

$$\begin{aligned} V\left(\sum_{i=1}^n w_{is} Y_i\right) &= V\left(\sum_{i=1}^N w'_{is} Y_i\right) \\ &= \sum_{i=1}^N Y_i^2 V(w'_{is}) + 2 \sum_{i=1}^N \sum_{j=1}^N Y_i Y_j \text{Cov}(w'_{is}, w'_{js}), \\ v\left(\sum_{i=1}^n w_{is} Y_i\right) &= v\left(\sum_{i=1}^N w'_{is} Y_i\right) \\ &= \sum_{i=1}^N a'_{is} Y_i^2 + 2 \sum_{i=1}^N \sum_{j=1}^N b'_{ijs} Y_i Y_j. \end{aligned}$$

根据定理的条件,

$$E \left[v \left(\sum_{i=1}^n w_{is} Y_i \right) \right] = V \left[\sum_{i=1}^N w'_{is} Y_i \right].$$

因此

$$\begin{aligned} & \sum_{i=1}^N E(a'_{is}) Y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N E(b'_{ijs}) Y_i Y_j \\ &= \sum_{i=1}^N V(w'_{is}) Y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N \text{Cov}(w'_{is}, w'_{js}) Y_i Y_j. \end{aligned}$$

从而必有

$$E(a'_{is}) = V(w'_{is}),$$

$$\begin{aligned} & E \left[v \left(\sum_{i=1}^n w_{is} \hat{P}_i \right) \right] \\ &= E_1 E_2 \left[\sum_{i=1}^N a'_{is} \hat{P}_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N b'_{ijs} \hat{P}_i \hat{P}_j \right] + E_1 E_2 \left[\sum_{i=1}^N w'_{is} \hat{\sigma}_{2i}^2 \right] \\ &= E_1 \left[\sum_{i=1}^N a'_{is} (Y_i^2 + \sigma_{2i}^2) \right] + 2 E_1 \left[\sum_{i=1}^N \sum_{j>i}^N b'_{ijs} Y_i Y_j \right] \\ &\quad + E_1 \left[\sum_{i=1}^N w'_{is} \sigma_{2i}^2 \right] \\ &= E_1 \left[\sum_{i=1}^N a'_{is} Y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N b'_{ijs} Y_i Y_j \right] \\ &\quad + \sum_{i=1}^N [E_1(a'_{is}) + E_1(w'_{is})] \sigma_{2i}^2. \end{aligned} \tag{7.67}$$

因为

$$\begin{aligned} E_1(a'_{is}) &= V_1(w'_{is}), \\ 1 &= E_1(w'_{is}) = [E_1(w'_{is})]^2, \end{aligned}$$

$$\begin{aligned} \text{故} \quad E \left[v \left(\sum_{i=1}^n w_{is} \hat{P}_i \right) \right] &= V \left(\sum_{i=1}^n w_{is} Y_i \right) + \sum_{i=1}^N E_1(w'_{is}^2) \sigma_{2i}^2 \\ &= V \left(\sum_{i=1}^n w_{is} \hat{P}_i \right) = V(\hat{P}). \quad \blacksquare \end{aligned}$$

推论 在定理 7.5 的条件下,

$$v' \left(\sum_{i=1}^n w_{is} \hat{P}_i \right) = \sum_{i=1}^n a_{is} \hat{P}_i^2 + 2 \sum_{i=1}^n \sum_{j>i}^n b_{ijs} \hat{P}_i \hat{P}_j + \sum_{i=1}^n (w_{is}^2 - a_{is}) \hat{\sigma}_{2i}^2 \tag{7.68}$$

也是 $V(\hat{P}) = V \left(\sum_{i=1}^n w_{is} \hat{P}_i \right)$ 的无偏估计.

证明 对 $v' \left(\sum_{i=1}^n w_{is} \hat{P}_i \right)$ 取均值, 根据 (7.67) 式有

$$\begin{aligned} E \left[v' \left(\sum_{i=1}^n w_{is} \hat{P}_i \right) \right] &= E_1 \left[\sum_{i=1}^N a'_{is} Y_i^2 + 2 \sum_{i=1}^N \sum_{j>i}^N b'_{ijs} Y_i Y_j \right] \\ &\quad + \sum_{i=1}^N E_1(w'_{is}^2) \sigma_{2i}^2 \end{aligned}$$

$$\begin{aligned}
 &= V\left(\sum_{i=1}^N w_{is} Y_i\right) + \sum_{i=1}^N E_1(w'_{is}) \sigma_{2i}^2 \\
 &= V\left(\sum_{i=1}^N w_{is} \hat{Y}_i\right) = V(\hat{Y}). \blacksquare
 \end{aligned}$$

而且这里可以允许 σ_{2i}^2 依赖于样本中的其他单元, 即可记成 σ_{2is}^2 .

定理 7.4 给出了作为线性估计 (7.60) 是无偏的条件, 而这个条件是很容易验证的. 该定理同时给出了这种估计的方差表达式. 而利用定理 7.5 及其推论容易构造方差的无偏估计. 规则是在第一阶抽样 $V\left(\sum_{i=1}^N w_{is} Y_i\right)$ 的某个无偏估计 (是一个二次型) 公式中用 \hat{Y}_i 代替 Y_i , 再加上一个有关第二阶抽样的附加项 $\sum_{i=1}^N w_{is} \hat{\sigma}_{2i}^2$ 或 $\sum_{i=1}^N (w_{is}^2 - a_{is}) \hat{\sigma}_{2i}^2$ 即可. 其中 $\hat{\sigma}_{2i}^2$ 应是 $V_2(\hat{Y}_i) = \sigma_{2i}^2$ 的无偏估计.

我们在以下两段中详细说明这两个定理的应用.

7.4.3 按简单随机抽样抽取初级单元

若第一阶抽样是按简单随机抽样抽取的, 此时第二阶抽样通常也是按简单随机抽样抽取. 考虑简单线性估计:

$$\hat{Y}_u = \frac{N}{n} \sum_{i=1}^n M_i \bar{y}_i = \frac{N}{n} \sum_{i=1}^n \hat{Y}_i, \quad (7.69)$$

此时

$$w_{is} = \frac{N}{n}. \quad (7.70)$$

因为
$$E(w'_{is}) = \frac{n}{N} \frac{N}{n} = 1,$$

所以
$$E(w'_{is}^2) = \frac{n}{N} \left(\frac{N}{n}\right)^2 = \frac{N}{n} \quad (i = 1, 2, \dots, N).$$

故根据定理 7.4, \hat{Y}_u 是无偏的, 且

$$\begin{aligned}
 V(\hat{Y}_u) &= V\left(\frac{N}{n} \sum_{i=1}^n Y_i\right) + \sum_{i=1}^N \frac{N}{n} \sigma_{2i}^2 \\
 &= \frac{N^2(1-f_1)}{n} \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N-1} + \frac{N}{n} \sum_{i=1}^N \frac{M_i^2(1-f_{2i})S_{2i}^2}{m_i}.
 \end{aligned} \quad (7.71)$$

其中
$$f_{2i} = \frac{m_i}{M_i}.$$

为寻求 $V(\hat{Y}_u)$ 的一个无偏估计, 我们注意到若令

$$\bar{Y}_u = \frac{1}{n} \sum_{i=1}^n Y_i,$$

则 $\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_u)^2$ 是 $\frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$ 的一个无偏估计, 又 $s_{2i}^2 = \frac{1}{m_i-1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2$ 是 $S_{2i}^2 = \frac{1}{M_i-1} \sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_i)^2$ 的一个无偏估计, 因此根据定理 7.5, 知

$$v(\hat{Y}_u) = \frac{N^2(1-f_1)}{n} \frac{\sum_{i=1}^n (\hat{Y}_i - \hat{\bar{Y}}_u)^2}{n-1} + \frac{N}{n} \sum_{i=1}^n \frac{M_i^2(1-f_{2i})s_{2i}^2}{m_i} \quad (7.72)$$

是 $V(\hat{Y}_u)$ 的一个无偏估计, 式中

$$\hat{\bar{Y}}_u = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i. \quad (7.73)$$

若将(7.69)式中的 \hat{Y}_u 改成为

$$\hat{Y}_u = \frac{N}{n} \sum_{i=1}^n \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij},$$

则容易看出, 当

$$f_{2i} = \frac{m_i}{M_i} = f_2 \quad (7.74)$$

为常数时, \hat{Y}_u 是自加权的.

简单估计 \hat{Y}_u 虽然是无偏的, 但效果一般不好, 方差较大. 此时我们还可考虑比估计. 例如对大小 M_i 的比估计形式为:

$$\hat{Y}_R = M_0 \frac{\sum_{i=1}^n M_i \bar{y}_i}{\sum_{i=1}^n M_i} = M_0 \cdot \frac{\sum_{i=1}^n \hat{Y}_i}{\sum_{i=1}^n M_i} \triangleq M_0 \hat{\bar{Y}}_R. \quad (7.75)$$

由于它不是线性估计, 故不能应用定理 7.4 与 7.5, 而且它是有偏的. 但根据与定理 4.1 类似的思路, 可以证明它的一个近似方差估计为:

$$v(\hat{Y}_R) = \frac{N^2(1-f_1)}{n} \frac{\sum_{i=1}^n M_i^2 (\bar{y}_i - \hat{\bar{Y}}_R)^2}{n-1} + \frac{N}{n} \sum_{i=1}^n \frac{M_i^2(1-f_{2i})S_{2i}^2}{m_i}. \quad (7.76)$$

7.4.4 按不放回不等概率抽取初级单元

设初级单元是按不放回不等概率抽取的, π_i, π_{ij} 是包含概率, 则此时总体总和 Y 的二阶估计可采用以下形式的 Horvitz-Thompson 估计:

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{M_i \bar{y}_i}{\pi_i} = \sum_{i=1}^n \frac{\hat{Y}_i}{\pi_i}. \quad (7.77)$$

此时 $w'_{is} = \begin{cases} \frac{1}{\pi_i}, & \text{若单元 } i \text{ 入样,} \\ 0, & \text{否则.} \end{cases}$

对固定的 i , $E_1(w'_{is}) = \pi_i \cdot \frac{1}{\pi_i} = 1$,

$$E_1(w'^2_{is}) = \pi_i \cdot \frac{1}{\pi_i^2} = \frac{1}{\pi_i}.$$

故 \hat{Y}_{HT} 是无偏的, 且

$$V(\hat{Y}_{HT}) = V\left[\sum_{i=1}^n \frac{Y_i}{\pi_i}\right] + \sum_{i=1}^n \frac{\sigma_{2i}^2}{\pi_i}.$$

根据定理 5.2, 当 n 固定时,

$$V\left[\sum_{i=1}^n \frac{Y_i}{\pi_i}\right] = \sum_{i=1}^n \sum_{j>i}^n (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2,$$

因此(若不限定第二阶抽样形式)

$$V(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i=1}^n \frac{\sigma_{2i}^2}{\pi_i}. \quad (7.78)$$

由于在单阶抽样中, Yates-Grundy Sen 估计量

$$v_{YGS} = \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j}\right)^2$$

是(7.78)式前一项的无偏估计. 因此若 $\hat{\sigma}_{2i}^2$ 是 σ_{2i}^2 的无偏估计, 根据定理 7.5, 按(7.66)式,

$$v(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j}\right)^2 + \sum_{i=1}^n \frac{\hat{\sigma}_{2i}^2}{\pi_i} \quad (7.79)$$

是 $V(\hat{Y}_{HT})$ 的一个无偏估计. 而按定理 7.5 的推论

$$\begin{aligned} v'(\hat{Y}_{HT}) &= \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j}\right)^2 \\ &\quad + \sum_{i=1}^n \sum_{j>i}^n \left(\frac{1}{n-1} - \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right) \left(\frac{\hat{\sigma}_{2i}^2}{\pi_i^2} + \frac{\hat{\sigma}_{2j}^2}{\pi_j^2}\right) \end{aligned} \quad (7.80)$$

也是 $V(\hat{Y}_{HT})$ 的一个无偏估计. 由于从(7.68)式到(7.80)式不很明显, 下面 we 进行直接验证.

$$\begin{aligned} &E\left[\sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j}\right)^2\right] \\ &= E_1\left\{\sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} E_2\left[\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right) + \left(\frac{\hat{Y}_i}{\pi_i} - \frac{Y_i}{\pi_i}\right)\right]\right\} \end{aligned}$$

$$\begin{aligned}
& - \left(\frac{\hat{P}_j}{\pi_j} - \frac{Y_j}{\pi_j} \right) \Big]^2 \Big\} \\
& = E_1 \left\{ \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left[E_2 \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 + E_2 \left(\frac{\hat{P}_i}{\pi_i} - \frac{Y_i}{\pi_i} \right)^2 \right. \right. \\
& \quad \left. \left. + E_2 \left(\frac{\hat{P}_j}{\pi_j} - \frac{Y_j}{\pi_j} \right)^2 \right] \right\} \\
& = \sum_{i=1}^N \sum_{j>i}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 \\
& \quad + E_1 \left[\sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\sigma_{2i}^2}{\pi_i^2} + \frac{\sigma_{2j}^2}{\pi_j^2} \right) \right] \\
& = V(\hat{P}_{HT}) - \sum_{i=1}^N \frac{\sigma_{2i}^2}{\pi_i} + E_1 \left[\sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\sigma_{2i}^2}{\pi_i^2} + \frac{\sigma_{2j}^2}{\pi_j^2} \right) \right] \\
& = V(\hat{P}_{HT}) + E \left[\sum_{i=1}^n \frac{\hat{\sigma}_{2i}^2}{\pi_i^2} - \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{\sigma}_{2i}^2}{\pi_i^2} + \frac{\hat{\sigma}_{2j}^2}{\pi_j^2} \right) \right].
\end{aligned}$$

所以

$$\sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{P}_i}{\pi_i} - \frac{\hat{P}_j}{\pi_j} \right)^2 + \sum_{i=1}^n \frac{\hat{\sigma}_{2i}^2}{\pi_i^2} - \sum_{i=1}^n \sum_{j>i}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{\sigma}_{2i}^2}{\pi_i^2} + \frac{\hat{\sigma}_{2j}^2}{\pi_j^2} \right)$$

是 \hat{P}_{HT} 的无偏估计, 注意到

$$\sum_{i=1}^n \frac{\hat{\sigma}_{2i}^2}{\pi_i^2} = \frac{1}{n-1} \sum_{i=1}^n \sum_{j>i}^n \left(\frac{\hat{\sigma}_{2i}^2}{\pi_i^2} + \frac{\hat{\sigma}_{2j}^2}{\pi_j^2} \right), \quad (7.81)$$

故 $\psi'(\hat{P}_{HT})$ 是 \hat{P}_{HT} 的无偏估计.

作为例子, 下面列出第一阶抽样分别按 Brewer-Durbin、Yates-Grundy 与 Rao-Hartley-Cochran 方法的二阶估计量及相应的无偏方差估计.

1) Brewer 或 Durbin 方法

$$\hat{P}_B = \frac{\hat{P}_1}{\pi_1} + \frac{\hat{P}_2}{\pi_2} = \frac{1}{2} \left(\frac{M_1 \bar{y}_1}{z_1} + \frac{M_2 \bar{y}_2}{z_2} \right), \quad (7.82)$$

$$\psi'(\hat{P}_B) = \frac{\pi_1 \pi_2 - \pi_{12}}{\pi_{12}} \left(\frac{\hat{P}_1}{\pi_1} - \frac{\hat{P}_2}{\pi_2} \right)^2 + \sum_{i=1}^2 \frac{\hat{\sigma}_{2i}^2}{\pi_i}, \quad (7.83)$$

$$\psi'(\hat{P}_B) = \frac{\pi_1 \pi_2 - \pi_{12}}{\pi_{12}} \left(\frac{\hat{P}_1}{\pi_1} - \frac{\hat{P}_2}{\pi_2} \right)^2 + \left(1 - \frac{\pi_1 \pi_2 - \pi_{12}}{\pi_{12}} \right) \left(\frac{\hat{\sigma}_{21}^2}{\pi_1^2} + \frac{\hat{\sigma}_{22}^2}{\pi_2^2} \right) \quad (7.84)$$

其中 $\pi_i = 2z_i$, 而 π_{12} 是由 (5.30) 式给出的.

2) $n=2$ 时的 Yates-Grundy 逐个抽取法, 用 Murthy 估计量

$$\hat{P}_M = \frac{1}{2} \frac{1}{z_1 z_2} \left[(1-z_2) \frac{\hat{P}_1}{z_1} + (1-z_1) \frac{\hat{P}_2}{z_2} \right], \quad (7.85)$$

$$v(\hat{P}_M) = \frac{(1-z_1)(1-z_2)(1-z_1-z_2)}{(2-z_1-z_2)^2} \left(\frac{\hat{P}_1}{z_1} - \frac{\hat{P}_2}{z_2} \right)^2 + \sum_{i=1}^2 \frac{\hat{\sigma}_{2i}^2}{\pi_i}, \quad (7.86)$$

$$\begin{aligned} v'(\hat{P}_M) = & \frac{(1-z_1)(1-z_2)(1-z_1-z_2)}{(2-z_1-z_2)^2} \left(\frac{\hat{P}_1}{z_1} - \frac{\hat{P}_2}{z_2} \right)^2 \\ & - \frac{(1-z_1)(1-z_2)(1-z_1-z_2)}{(2-z_1-z_2)^2} \left(\frac{\hat{\sigma}_{21}^2}{z_1^2} + \frac{\hat{\sigma}_{22}^2}{z_2^2} \right) \\ & + \frac{(1-z_2)^2}{(2-z_1-z_2)^2} \cdot \frac{\hat{\sigma}_{21}^2}{z_1^2} + \frac{(1-z_1)^2}{(2-z_1-z_2)^2} \cdot \frac{\hat{\sigma}_{22}^2}{z_2^2}. \end{aligned} \quad (7.87)$$

3) Rao-Hartley Cochran 抽样及估计量

$$\hat{P}_{RHC} = \sum_{g=1}^n Z_g^* \frac{\hat{P}_g}{z_g}, \quad (7.88)$$

$$v(\hat{P}_{RHC}) = \frac{\sum_{g=1}^n N_g^2 - N}{N^2 \sum_{g=1}^n N_g^2} \sum_{g=1}^n Z_g^* \left(\frac{\hat{P}_g}{z_g} - \hat{P}_{RHC} \right)^2 + \sum_{g=1}^n Z_g^* \frac{\hat{\sigma}_{2g}^2}{z_g^2}, \quad (7.89)$$

$$\begin{aligned} v'(\hat{P}_{RHC}) = & \frac{\sum_{g=1}^n N_g^2 - N}{N^2 - \sum_{g=1}^n N_g^2} \sum_{g=1}^n Z_g^* \left(\frac{\hat{P}_g}{z_g} - \hat{P}_{RHC} \right)^2 \\ & - \left\{ (1 - Z_g^*) \frac{\hat{\sigma}_{2g}^2}{z_g^2} + \sum_{u=1}^n \sum_{h>g} Z_g^{*2} \frac{\hat{\sigma}_{2h}^2}{z_h^2} \right\} + \sum_{g=1}^n Z_g^{*2} \frac{\hat{\sigma}_{2g}^2}{z_g^2}. \end{aligned} \quad (7.90)$$

§ 7.5 三阶及多阶抽样

7.5.1 各级单元大小相等时的三阶抽样

前几节都是对二阶抽样情形讨论的。用类似方法可以将这些结果推广到三阶或更高阶的抽样。例如在各级单元大小都相等情况的三阶抽样，有类似 § 7.2 节的结果。

设总体中含有 N 个一级单元，每个包含 M 个二级单元，而每个二级单元又包含 K 个三级单元。设三阶抽样的各阶样本量分别为 n, m 与 k ，我们引进以下记号：

Y_{iju} 为第 i 个一级单元，第 j 个二级单元，第 u 个三级单元的观测值； y_{iju} 为相应的样本值；

$$\begin{aligned}
\bar{Y}_{ij} &= \frac{1}{K} \sum_{u=1}^K Y_{iju}, & \bar{y}_{ij} &= \frac{1}{k} \sum_{u=1}^k y_{iju}, \\
\bar{Y}_i &= \frac{1}{MK} \sum_{j=1}^M \sum_{u=1}^K Y_{iju}, & \bar{y}_i &= \frac{1}{mk} \sum_{j=1}^m \sum_{u=1}^k y_{iju}, \\
\bar{\bar{Y}} &= \frac{1}{NMK} \sum_{i=1}^N \sum_{j=1}^M \sum_{u=1}^K Y_{iju}, & \bar{\bar{y}} &= \frac{1}{nmk} \sum_{i=1}^n \sum_{j=1}^m \sum_{u=1}^k y_{iju}, \\
S_1^2 &= \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{\bar{Y}})^2, & s_1^2 &= \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \bar{\bar{y}})^2, \\
S_2^2 &= \frac{1}{N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2, \\
s_2^2 &= \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2, \\
S_3^2 &= \frac{1}{NM(K-1)} \sum_{i=1}^N \sum_{j=1}^M \sum_{u=1}^K (Y_{iju} - Y_{ij})^2, \\
s_3^2 &= \frac{1}{nm(k-1)} \sum_{i=1}^n \sum_{j=1}^m \sum_{u=1}^k (y_{iju} - \bar{y}_{ij})^2.
\end{aligned}$$

在此情形,有如下定理:

定理 7.6 若在三阶抽样中,每阶抽样都是简单随机的,则

1)

$$E(\bar{\bar{y}}) = \bar{\bar{Y}}, \quad (7.91)$$

2)

$$V(\bar{\bar{y}}) = \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{nm} S_2^2 + \frac{1-f_3}{nmk} S_3^2, \quad (7.92)$$

其中

$$f_1 = \frac{n}{N}, \quad f_2 = \frac{m}{M}, \quad f_3 = \frac{k}{K}.$$

3)

$$v(\bar{\bar{y}}) = \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-f_2)}{nm} s_2^2 + \frac{f_1 f_2 (1-f_3)}{nmk} s_3^2 \quad (7.93)$$

是 $V(\bar{\bar{y}})$ 的无偏估计.

本定理的证明完全类似于定理 7.1 与定理 7.2 的证明. 这里从略. 在证明(7.93)过程中可得到:

$$E(s_1^2) = S_1^2 + \frac{1-f_2}{m} S_2^2 + \frac{1-f_3}{mk} S_3^2, \quad (7.94)$$

$$E(s_2^2) = S_2^2 + \frac{1-f_3}{k} S_3^2, \quad (7.95)$$

$$E(s_3^2) = S_3^2. \quad (7.96)$$

由此可知, s_3^2 是 S_3^2 的无偏估计, 而 s_2^2 、 s_1^2 不是 S_2^2 与 S_1^2 的无偏估计.

但根据上面的表达式, 不难构造它们的无偏估计.

从定理 7.6 可知, 在三阶抽样中, 一般而言, 第一阶抽样的方差是最主要的, 第二阶抽样的方差次之, 第三阶抽样的方差已相当小, 通常可忽略不计. 这个规律也适用于更高阶的抽样. 事实上, 对于高阶抽样, 一般仅需计算前两阶抽样的方差即可.

若抽样的总费用函数具有以下简单的线性形式:

$$C = c_0 + c_1 n + c_2 nm + c_3 nmk, \quad (7.97)$$

则 k, m 的最优值(使总方差达到最小)分别为:

$$k_{\text{opt}} = \frac{S_3}{\sqrt{S_2^2 - S_3^2/K}} \sqrt{\frac{c_2}{c_3}}, \quad m_{\text{opt}} = \frac{\sqrt{S_2^2 - S_3^2/K}}{\sqrt{S_1^2 - S_2^2/M}} \sqrt{\frac{c_1}{c_2}}. \quad (7.98)$$

根据以上公式的结构, 不难得出更高阶抽样的相应公式.

7.5.2 多阶抽样中不等概率抽样的应用

对于一般情形的多阶抽样, 即各级单元大小不等的情形, 普遍采用不等概率抽样, 因为它不仅效率高(方差小), 而且若各阶抽样的概率选择得合理, 还可以大大简化计算. 仍以三阶抽样为例, 有如下定理:

定理 7.7 在三阶抽样中, 设每一阶抽样都是按多项抽样抽取的, 各阶样本量分别为 n, m, k , 抽样概率分别为 Z_i, Z_{ij}, Z_{iju} ($\sum_{i=1}^N Z_i = 1, \sum_{j=1}^{M_i} Z_{ij} = 1, \sum_{u=1}^{K_{ij}} Z_{iju} = 1, i=1, 2, \dots, N; j=1, 2, \dots, M_i; u=1, 2, \dots, K_{ij}$), 记 y_{iju} 为样本观测值, 则总体总和 Y 的以下估计量

$$\hat{P} = \frac{1}{nmk} \sum_{i=1}^n \frac{1}{Z_i} \sum_{j=1}^m \frac{1}{Z_{ij}} \sum_{u=1}^k \frac{y_{iju}}{Z_{iju}} \triangleq \frac{1}{n} \sum_{i=1}^n \hat{P}_i \quad (7.99)$$

是无偏的, 且

$$\begin{aligned} V(\hat{P}) &= \frac{1}{n} \left(\sum_{i=1}^N \frac{Y_i^2}{Z_i} - Y^2 \right) + \frac{1}{nm} \sum_{i=1}^N \frac{1}{Z_i} \left(\sum_{j=1}^{M_i} \frac{Y_{ij}^2}{Z_{ij}} - Y_i^2 \right) \\ &\quad + \frac{1}{nmk} \sum_{i=1}^N \frac{1}{Z_i} \left[\sum_{j=1}^{M_i} \frac{1}{Z_{ij}} \left(\sum_{u=1}^{K_{ij}} \frac{Y_{iju}^2}{Z_{iju}} - Y_{ij}^2 \right) \right], \end{aligned} \quad (7.100)$$

而

$$v(\hat{P}) = \frac{1}{n(n-1)} \sum_{i=1}^n (\hat{P}_i - \hat{P})^2 = \frac{1}{n(n-1)} \left[\sum_{i=1}^n \hat{P}_i^2 - n\hat{P}^2 \right] \quad (7.101)$$

是 $V(\hat{P})$ 的一个无偏估计, 其中

$$\hat{P}_i = \frac{1}{Z_i m} \sum_{j=1}^m \frac{1}{Z_{ij}} \left(\frac{1}{k} \sum_{u=1}^k \frac{y_{iju}}{Z_{iju}} \right) \quad (i=1, 2, \dots, n).$$

证明 反复应用证明定理 5.1 的方法与结论以及 (7.3) 与 (7.4) 式, 即有

$$\begin{aligned}
 E(\hat{P}) &= E_1 E_2 E_3(\hat{P}) \\
 &= E_1 \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{z_i} E_2 \left[\frac{1}{m} \sum_{j=1}^m \frac{1}{z_{ij}} E_3 \left(\frac{1}{k} \sum_{u=1}^k \frac{y_{iju}}{z_{iju}} \right) \right] \right\} \\
 &= E_1 \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{z_i} E_2 \left[\frac{1}{m} \sum_{j=1}^m \frac{Y_{ij}}{z_{ij}} \right] \right\} \\
 &= E_1 \left\{ \frac{1}{n} \sum_{i=1}^n \frac{Y_i}{z_i} \right\} = Y, \\
 V(\hat{P}) &= V_1 E_2 E_3(\hat{P}) + E_1 V_2 E_3(\hat{P}) + E_1 E_2 V_3(\hat{P}), \\
 V_1 E_2 E_3(\hat{P}) &= V_1 \left[\frac{1}{n} \sum_{i=1}^n \frac{Y_i}{z_i} \right] \\
 &= \frac{1}{n} \sum_{i=1}^N Z_i \left(\frac{Y_i}{Z_i} - Y \right)^2 = \frac{1}{n} \left[\sum_{i=1}^N \frac{Y_i^2}{Z_i} - Y^2 \right], \\
 E_1 V_2 E_3(\hat{P}) &= \frac{1}{n^2} E_1 \left\{ \sum_{i=1}^n \frac{1}{z_i^2} V_2 \left[\frac{1}{m} \sum_{j=1}^m \frac{Y_{ij}}{z_{ij}} \right] \right\} \\
 &= \frac{1}{n} E_1 \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{z_i^2} \cdot \frac{1}{m} \left(\sum_{j=1}^{M_i} \frac{Y_{ij}^2}{Z_{ij}} - Y_i^2 \right) \right\} \\
 &= \frac{1}{nm} \sum_{i=1}^N \frac{1}{Z_i} \left(\sum_{j=1}^{M_i} \frac{Y_{ij}^2}{Z_{ij}} - Y_i^2 \right), \\
 E_1 E_2 V_3(\hat{P}) &= \frac{1}{nm} E_1 \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{z_i^2} \cdot \frac{1}{m} E_2 \left[\sum_{j=1}^m \frac{1}{z_{ij}^2} \right. \right. \\
 &\quad \left. \left. \times \frac{1}{k} \left(\sum_{u=1}^{K_{ij}} \frac{Y_{iju}^2}{Z_{iju}} - Y_{ij}^2 \right) \right] \right\} \\
 &= \frac{1}{nmk} E_1 \left\{ \frac{1}{n} \sum_{i=1}^n \frac{1}{z_i^2} \sum_{j=1}^{M_i} \frac{1}{Z_{ij}} \left(\sum_{u=1}^{K_{ij}} \frac{Y_{iju}^2}{Z_{iju}} - Y_{ij}^2 \right) \right\} \\
 &= \frac{1}{nmk} \sum_{i=1}^N \frac{1}{Z_i} \left[\sum_{j=1}^{M_i} \frac{1}{Z_{ij}} \left(\sum_{u=1}^{K_{ij}} \frac{Y_{iju}^2}{Z_{iju}} - Y_{ij}^2 \right) \right].
 \end{aligned}$$

因而 (7.100) 式成立. 为证明 (7.101) 式, 我们需注意到第一阶抽样是放回的, 故 n 个无偏估计量

$$\hat{P}_i = \frac{1}{z_i} \frac{1}{m} \sum_{j=1}^m \frac{1}{z_{ij}} \left(\frac{1}{k} \sum_{u=1}^k \frac{y_{iju}}{z_{iju}} \right)$$

是相互独立的, 且具有相同方差, 因而

$$\frac{1}{n(n-1)} \sum_{i=1}^n (\hat{P}_i - \hat{P})^2 = \frac{1}{n(n-1)} \left[\sum_{i=1}^n \hat{P}_i^2 - n\hat{P}^2 \right]$$

是 \hat{P}_i 的平均数 $\hat{P} = \frac{1}{n} \sum_{i=1}^n \hat{P}_i$ 的方差 $V(\hat{P})$ 的一个无偏估计. ■

推论 在三阶抽样中, 若前两阶抽样都是 PPS 抽样, 最后一阶是按

等概率抽取的, 又各阶样本量对不同单元都等于常数, 则所得的样本是自加权的.

事实上, 此时有

$$Z_i = \frac{M_i}{M_0}, \quad Z_{ij} = \frac{K_{ij}}{M_i}, \quad Z_{iju} = \frac{1}{K_{ij}}.$$

其中 $M_0 = \sum_{i=1}^N M_i = \sum_{i=1}^N \sum_{j=1}^{M_i} K_{ij}$ 是总体中所有三级单元的总数. 根据定理 7.7, 此时总体总和 Y 的估计可写成

$$\hat{Y} = \frac{M_0}{nmk} \sum_{i=1}^n \sum_{j=1}^m \sum_{u=1}^k y_{iju} \triangleq M_0 \bar{y}. \quad (7.102)$$

其中

$$\bar{y} = \frac{1}{nmk} \sum_{i=1}^n \sum_{j=1}^m \sum_{u=1}^k y_{iju} \quad (7.103)$$

是样本观测值按三级单元的简单平均数. 此时方差估计也有极简单的形式:

$$v(\hat{Y}) = \frac{M_0^2}{n(n-1)} \sum_{i=1}^n (\bar{y}_i - \bar{y})^2. \quad (7.104)$$

其中

$$\bar{y}_i = \frac{1}{mk} \sum_{j=1}^m \sum_{u=1}^k y_{iju} \quad (i = 1, 2, \dots, n). \quad (7.105)$$

若最后一阶抽样是(不放回)简单随机抽样, 则上述结论仍然成立, 不过此时理论方差 $V(\hat{Y})$ 比第三阶抽样是有放回等概率抽样小.

从上面的讨论不难得到更高阶抽样的一般处理方法. 如果在多阶抽样中, 前几阶都是按有放回的 PPS 抽样, 最后一阶为等概率抽样, 那么所得的样本是自加权的, 估计量及其方差估计都有简单的形式.

第 8 章

系统抽样

§ 8.1 一般描述

8.1.1 定义及实施方法

定义 8.1 设总体中的 N 个单元按一定顺序 (随机的或按某种规律排列), 编号为 $1, 2, \dots, N$, 采取如下方法从总体中抽取一个样本量为 n 的样本: 先抽取一个或一组随机数字作为起始单元的编号, 然后按一个确定的规则抽取其他单元. 这种抽样称为系统抽样 (systematic sampling).

系统抽样中一种最简单的方法是在抽取起始单元的编号后, 按一确定间距 k (k 为最接近于 N/n 的整数), 逐个抽取样本单元. 这种系统抽样也称为等距抽样. 其中 k 称为抽样间距 (sampling interval), 具体地说, 先在 1 至 k 之间随机地抽取一个整数 i , 以它作为起始单元的编号, 则整个样本是由以下编号的单元组成的.

$$i + (j-1)k \quad (j=1, 2, \dots, n).$$

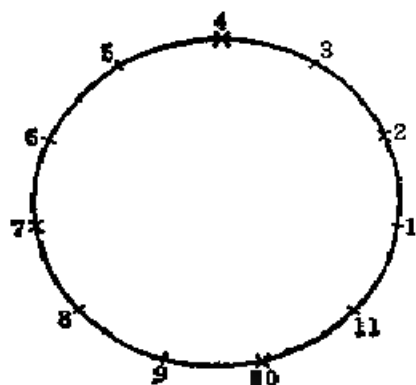


图 8.1 圆形系统抽样示意图

由于 N 不一定是 k 的整数倍, 所以按上述方法得到的系统样本的样本量可为 $\left[\frac{N}{k} \right]$ 或 $\left[\frac{N}{k} \right] + 1$. 为了避免这种样本量不能确定的情况, Lahiri (1952) 提出如下称为圆形系统抽样的方法. 将 N 个总体单元排列成一个圆, 首尾相接. 从 1 到 N 中抽取一个随机整数 i 作为初始单元, 然后每间隔 k 抽取一个单元 (k 仍为最接近于 N/n 的整数), 直至抽足 n 个单元为止. 按此方法, 可以保证样本量 n

不变。不过此时首尾两个样本单元的间隔可能小于 k , 也可能大于 k 。例如图 8.1 中, $N=11$, $k=3$, $n=4$, $i=4$, 首尾两个样本单元的间隔是 2。

从上述的实施方法可以看出, 在系统抽样过程中, 一旦起始单元确定了, 整个样本就完全确定了, 这是系统抽样有别于其他抽样的一个特点。

另外, 我们注意到, 当 $N=nk$ 时, 在上述两种实施方法中, 无论按哪一种方法, 总体中每个单元的入样概率都相等, 从而是一种等概率抽样。当 $N \neq nk$ 时, 按第一种方法每一个单元的入样概率依赖于初始值 i 。对不同的 i , 稍有不同。以下为了处理方便, 我们假定 N 总是 n 的整数倍。在实际问题中, 若 n 比较大, 例如 $n \geq 50$, 就可以不考虑 N/n 不是整数所带来的问题。

除了上述两种最简单的系统抽样, 即等距抽样外, 还有几种其他类型的系统抽样, 包含不等概率系统抽样, 将分别在 § 8.4 与 § 8.5 中作介绍。

8.1.2 系统抽样与整群抽样和分层抽样的关系

系统抽样可以看成是一种特殊的整群抽样, 也可以看成是一种分层抽样。为了看清其中的关系, 我们以一般的等距抽样为例, 将总体中的 $N(=nk)$ 个单元按 k 个一组排列成表 8.1 形式, 共有 k 行 n 列, 并以行、列号将单元进行重新编号。

表 8.1 系统抽样总体单元按群(行)、层(列)的排列

	1	2	...	j	...	n	行平均
1	Y_{11}	Y_{12}	...	Y_{1j}	...	Y_{1n}	$\bar{Y}_{1.}$
2	Y_{21}	Y_{22}	...	Y_{2j}	...	Y_{2n}	$\bar{Y}_{2.}$
\vdots	\vdots	\vdots		\vdots		\vdots	
i	Y_{i1}	Y_{i2}	...	Y_{ij}	...	Y_{in}	$\bar{Y}_{i.}$
\vdots	\vdots	\vdots		\vdots		\vdots	
k	Y_{k1}	Y_{k2}	...	Y_{kj}	...	Y_{kn}	$\bar{Y}_{k.}$
列平均	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$...	$\bar{Y}_{.j}$...	$\bar{Y}_{.n}$	\bar{Y}

每一个系统样本都是由表 8.1 中的一行单元所组成的。如果将每一行单元看作为一个群(大小为 n), 则总体由 k 个群组成。由于初始单元 i

(即行号)是随机抽取的, 因此这种系统抽样可以看成是对群进行随机抽样的整群抽样。为了以后能直接采用整群抽样的某些结果, 表 8.2 列出了系统抽样与整群抽样参数间的对照。

表 8.2 系统抽样与整群抽样若干参数对照表

系 统 抽 样	N	n	k	1	\bar{Y}
整 群 抽 样	NM	M	N	n	$\bar{\bar{Y}}$

另一方面, 若将表 8.1 中的列看成为层, 则每个系统样本都包含每层中的一个单元, 因此系统抽样也是一种分层抽样。不过由于样本单元在层中的位置都是一样的, 因此它不是分层随机抽样。

8.1.3 系统抽样的优缺点

系统抽样是实际中最常用的抽样方法之一。这是因为它有突出的优点: 首先系统抽样简单易行, 容易确定样本单元。它避免了一般概率抽样的诸多麻烦, 甚至在某些场合不需要抽样框。它所需要的只是总体单元的排列顺序。例如若要对公路旁的树木进行病虫害调查, 确定每 20 棵树检查一棵, 只要在初始被检树确定后, 每隔 20 棵树检查一棵即行, 根本不需要在事先对公路旁的所有树木进行编号。另外一些情况更为简单, 例如为对某城市的机动车辆进行调查, 确定抽样比为 1%, 则可在 00~99 中随机抽取一个整数, 不妨设是 42, 则只要车辆牌照号末两位为 42 的都进行调查即可。系统抽样的第二个优点是样本单元在总体中分布比较均匀, 因此在通常情形, 系统样本一般具有对总体的较好代表性, 这也是它受到实际工作者欢迎的一个重要原因。最后我们还应指出: 如果抽样者对总体结构(主要是指单元指标与排列顺序的关系)有较多的了解并加以正确利用的话, 系统抽样可以大大提高精度。

不过系统抽样也有其突出的缺点: 如果抽样者缺乏经验, 对于某些总体, 例如单元指标随着排列顺序呈周期性变化的情形, 处理不好, 系统抽样的效果就会大大降低。另外由于许多在实际中表明是行之有效的系统抽样常常不是严格的概率抽样, 因此系统抽样的方差估计较为困难。

§ 8.2 等概率系统抽样(等距抽样)

本节首先讨论最简单的系统抽样,即等距抽样时总体均值 \bar{Y} 的估计问题. 为讨论方便起见,仍假定 $N = nk$, 在此情形,抽样是一种严格的概率抽样.

8.2.1 估计量

按表 8.1 的记号,设初始单元编号为 i , 则总体均值 \bar{Y} 的估计量取为系统样本的均值:

$$\bar{y}_{sy} = \bar{Y}_{i.} \triangleq \frac{1}{n} \sum_{j=1}^n Y_{ij}. \quad (8.1)$$

由于可能样本只有 k 个, 因此

$$E(\bar{y}_{sy}) = \frac{1}{k} \sum_{i=1}^k \bar{Y}_{i.} = \frac{1}{kn} \sum_{i=1}^k \sum_{j=1}^n Y_{ij} = \bar{Y}. \quad (8.2)$$

因而估计量是无偏的.

如果 N 不是 n 或 k 的整数倍, 则上述估计量是有偏的, 不过当 n 比较大时, 其偏倚不会很大.

8.2.2 估计量的方差——用样本(群)内方差表示

估计量 \bar{y}_{sy} 的方差有几种不同的表达形式. 第一种形式是用系统样本也即群内方差 S_{wsy}^2 来表示.

定理 8.1 系统抽样估计量 \bar{y}_{sy} 的方差可表达为:

$$V(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_{wsy}^2. \quad (8.3)$$

其中 S^2 为总体方差, 又

$$S_{wsy}^2 = \frac{1}{k(n-1)} \sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2 \quad (8.4)$$

是系统样本(群)内方差.

证明 将总体平方和按表 8.1 中的行(也即全部可能的系统样本, 或称为群)进行分解:

$$(N-1)S^2 = n \sum_{i=1}^k (\bar{Y}_{i.} - \bar{Y})^2 + \sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2,$$

另一方面, 根据定义, 有

$$V(\bar{y}_{sy}) = E(\bar{y}_{sy} - \bar{Y})^2 = \frac{1}{k} \sum_{i=1}^k (\bar{Y}_{..} - \bar{Y})^2,$$

因此
$$V(\bar{y}_{sy}) = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_{wsy}^2. \blacksquare$$

从本定理的结果可知, 系统样本内的方差愈大, 则估计量的方差愈小, 若将(8.3)式与简单随机抽样的方差公式(简单估计情形)作比较, 立即可得到系统抽样比简单随机抽样更为精确的条件是:

$$S_{wsy}^2 > S^2. \quad (8.5)$$

因此为了提高系统抽样的精度, 只要有可能, 将总体单元重新排列, 尽可能增大样本内的方差, 即可达到目的.

8.2.3 估计量的方差——用样本(群)内相关表示

完全等价的, 系统抽样估计量的方差也可用系统样本(群)内相关 ρ_{wsy} 来表示.

定理 8.2 系统抽样估计量 \bar{y}_{sy} 的方差可表达为:

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1)\rho_{wsy}]. \quad (8.6)$$

其中

$$\begin{aligned} \rho_{wsy} &= \frac{E(Y_{ij} - \bar{Y})(Y_{iu} - \bar{Y})}{E(Y_{ij} - \bar{Y})^2} \\ &= \frac{2}{(n-1)(N-1)S^2} \sum_{i=1}^k \sum_{j < u}^n (Y_{ij} - \bar{Y})(Y_{iu} - \bar{Y}) \end{aligned} \quad (8.7)$$

是系统样本(群)内相关.

证明 考虑到系统抽样是一种特殊的整群抽样(且群的大小都相等), 所以可直接利用整群抽样的结果, 根据定理 6.1, 按(6.14)式, 有

$$V(\bar{y}) = \frac{1-f}{n} \frac{NM}{M^2(N-1)} S^2 [1 + (M-1)\rho_c].$$

其中记号按标准整群抽样的形式, 由于所考虑的系统抽样是一种特殊的整群抽样, 其参数与标准的整群抽样有如表 8.2 的对应关系, 因此有

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \left(\frac{N-1}{N} \right) [1 + (n-1)\rho_{wsy}]. \blacksquare$$

从定理 8.2 可知, 系统样本(群)内相关愈大, 也即系统样本(群)内单元愈相似, 差别愈小, 则估计量的方差愈大, 这个结论与定理 8.1 的结论是一致的.

前面讨论的是系统抽样估计量的理论方差, 由于我们在抽样时, 实际

抽到的只是一个系统样本, 因此要给出 $V(\bar{y}_{sr})$ 的无偏估计是不可能的. 在 § 8.6 中我们专门来讨论方差估计问题.

8.2.4 数值例子

设有一个 $N=25$ 的人为总体, 按表 8.1 的形式排成 5 行 5 列, 如表 8.3 的左上部分. 对该总体抽取 $n=5$ 的一个样本, 我们来研究不同抽样方法的效果.

表 8.3 一个 $N=25$ 的人为总体及其方差分解

列 行	1	2	3	4	5	行 方 差
1	12	19	25	28	28	47.3
2	24	28	29	33	36	21.5
3	18	30	36	39	39	78.3
4	26	34	40	48	44	74.8
5	29	29	46	52	50	128.7
列方差	46.2	30.5	70.7	100.5	63.8	$S^2=103.86$

根据表 8.3, 有:

总体方差

$$S^2 = 103.86,$$

行(内)平均方差

$$S_r^2 = \frac{1}{5} (47.3 + 21.5 + \cdots + 128.7) = 70.12,$$

列(内)平均方差

$$S_c^2 = \frac{1}{5} (46.2 + 30.5 + \cdots + 63.8) = 63.34.$$

我们来比较几种不同的抽样方法, 每种 n 均为 5. 计算每种抽样估计量 \bar{y} 的方差.

1) 简单随机抽样

$$V_{na} = \frac{1-f}{n} S^2 = \frac{0.8}{5} \times 103.86 = 16.6176.$$

2) 以行为群(系统样本)的整群抽样或系统抽样

$$V_{\text{sys}} = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_r^2 = 43.6096.$$

3) 以列为群(系统样本)的整群抽样或系统抽样

$$V_{\text{cxy}} = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} S_c^2 = 49.0336.$$

4) 以行为层的分层随机抽样(每层抽1个单元)

$$V_{\text{rst}} = \frac{1-f}{n} S_r^2 = 11.2192.$$

5) 以列为层的分层随机抽样(每层抽1个单元)

$$V_{\text{cst}} = \frac{1-f}{n} S_c^2 = 10.1344.$$

从上述结果可以看到, 由于(平均)行方差与列方差均小于总体方差 S^2 , 因此本例中的系统抽样的效果不及简单随机抽样。至于对分层随机抽样, 正如预料的那样, 效果不仅优于系统抽样, 也优于简单随机抽样。

为了看出总体单元不同排列对系统抽样的影响, 我们将总体单元重新排列。表 8.4 是将单元按从小到大的顺序逐列排列, 而表 8.5 是按某种随机化程序将单元随机排列。分别观察以行为系统样本的系统抽样。

表 8.4 表 8.3 总体按大小顺序重新排列及其方差分解

列 \ 行	1	2	3	4	5	行 方 差
1	12	26	29	36	44	142.8
2	18	28	29	36	46	107.8
3	19	28	30	39	48	122.7
4	24	28	33	39	50	103.7
5	25	29	34	40	52	111.5
列方差	27.5	1.2	5.5	3.5	10.0	$S^2=103.86$

根据表 8.4, 平均行(内)方差为 117.7, 故系统抽样方差为:

$$V_{\text{sys}} = \frac{24}{25} \times 103.86 - \frac{5 \times 4}{25} \times 117.7 = 5.5456.$$

而按简单随机抽样的方差 V_{rd} 仍为 16.6176, 可见此时系统抽样的效果优于简单随机抽样。实际上, 将总体单元按大小顺序排列的目的就是为了增大系统样本内方差, 从而必然提高精度。

表 8.5 是将总体单元按随机置换重新排列而成, 平均行内方差为 102.06, 平均列内方差为 102.38, 均接近于总体方差, 因此无论以行还是以列为系统样本的系统抽样的方差都接近简单随机抽样的方差。

表 8.5 表 8.3 总体按随机顺序重新排列及其方差分解

列 行	1	2	3	4	5	行 方 差
1	39	28	40	36	29	31.3
2	19	18	29	26	46	127.3
3	30	36	33	52	50	102.2
4	12	39	44	28	34	152.8
5	25	48	28	24	29	96.7
列方差	106.5	120.2	48.7	131.2	96.3	$S^2=103.86$

§ 8.3 方差与总体单元排列顺序的关系

从上节结果可以看出, 系统抽样的精度不仅与总体方差有关, 也与样本(群)内方差有关。而这里的“群”完全是以单元的排列顺序确定的, 因此系统抽样的精度与总体单元的排列顺序有密切关系。本节详细讨论它们之间的关系, 分三种典型情况。

8.3.1 随 机 排 列

在许多情况下, 采用系统抽样主要是因为抽样方便。此时单元的排列多呈自然的顺序, 与其指标值无任何相关关系。这种排列称为随机排列, 也称按(与指标值)无关标识排列。典型的例子是当抽样单元为人时, 人员的排列是按姓氏笔划(或字母)顺序排列的情况。当抽样单元为各级行政单位时, 按其地址码排列或在其他单元情况按目录顺序排列等情形均可视为随机排列。

正如 8.2.4 段中的数值例子表明的那样, 当单元排列为随机时, 系统抽样与简单随机抽样有大致相同的效果, 也即从某种意义上而言, 两者的方差是相等的。不过我们注意到, 系统抽样的方差在很大程度上依赖于单

元的不同排列,而简单随机抽样对于固定的总体,方差是不变的.因此这里所说的某种意义是就平均意义而言的.而这里的“平均”又有两种解释.

第一种解释是将总体看作是固定的,正如迄今为止我们所一直理解的那样,它由 N 个确定的单元 $\{Y_1, Y_2, \dots, Y_N\}$ 组成.这 N 个单元有 $N!$ 种不同的排列,而每一种排列对应于一个按此作系统抽样的方差,所谓平均系指这 $N!$ 个方差的平均值.

定理 8.3 对固定的有限总体 $\{Y_1, Y_2, \dots, Y_N\}$, 以 V_{ran} 表示从中抽取样本量为 n 的简单随机样本估计量的方差, 以 V_{sy} 表示对某个确定的单元排列进行系统抽样估计量 (样本量皆为 n) 的方差, 则全部 $N!$ 种单元不同排列的 V_{sy} 的均值 $E(V_{\text{sy}})$ 满足

$$E(V_{\text{sy}}) = V_{\text{ran}}. \quad (8.8)$$

证明 对每一个确定的单元顺序, 仍按表 8.1 的形式排成 k 行 n 列 ($N = kn$) 的形式, 其单元指标值记为 Y_{ij} , 则

$$V_{\text{sy}} = \frac{1}{k} \sum_{i=1}^k (Y_{i1} - Y)^2 = \frac{1}{k} \sum_{i=1}^k \bar{Y}_i^2 - \bar{Y}^2.$$

其中 $\bar{Y}_i = \frac{1}{n} \sum_{j=1}^n Y_{ij}$. 若将对所有单元全部 $N!$ 种不同的排列求和记为 $\sum_A^{N!}$, 对取遍 N 个总体单元 Y_u 的求和记为 \sum_u^N , 对取遍 N 个总体单元中任意两个单元 Y_u, Y_v 的任何排列求和记为 $\sum_A^{N(N-1)}$ (为避免记号过于繁琐起见, 在后两种情形, 总体单元均只用一个下标), 则

$$\begin{aligned} E(V_{\text{sy}}) &= \frac{1}{N!} \sum_A^{N!} \left[\frac{1}{k} \sum_{i=1}^k \bar{Y}_i^2 - \bar{Y}^2 \right] \\ &= \frac{1}{N!} \left[\frac{1}{k} \sum_A^{N!} \sum_{i=1}^k \left(\frac{1}{n} \sum_{j=1}^n Y_{ij} \right)^2 - N! \bar{Y}^2 \right] \\ &= \frac{1}{N! kn^2} \sum_{i=1}^k \sum_A^{N!} \left[\sum_{j=1}^n Y_{ij}^2 + \sum_{j \neq l} Y_{ij} Y_{il} \right] - \bar{Y}^2 \\ &= \frac{1}{N! kn^2} \sum_{i=1}^k \left[\sum_{j=1}^n \sum_A^{N!} Y_{ij}^2 + \sum_{j \neq l} \sum_A^{N!} Y_{ij} Y_{il} \right] - \bar{Y}^2 \\ &= \frac{1}{N! kn^2} \sum_{i=1}^k \left[\sum_{j=1}^n (N-1)! \sum_u^N Y_u^2 + \sum_{j \neq l} (N-2)! \sum_A^{N(N-1)} Y_u Y_v \right] - \bar{Y}^2 \\ &= \frac{1}{N! kn^2} \sum_{i=1}^k \left[n(N-1)! \sum_u^N Y_u^2 + n(n-1)(N-2)! \sum_A^{N(N-1)} Y_u Y_v \right] - \bar{Y}^2 \\ &= \frac{1}{Nn} \sum_u^N Y_u^2 + \frac{n-1}{N(N-1)n} \sum_A^{N(N-1)} Y_u Y_v - \bar{Y}^2 \end{aligned}$$

$$\begin{aligned}
&= \left[\frac{1}{Nn} - \frac{n-1}{N(N-1)n} \right] \sum_{i=1}^N Y_u^2 + \frac{n-1}{N(N-1)n} \left(\sum_{i=1}^N Y_u \right)^2 - \bar{Y}^2 \\
&= \frac{N-n}{Nn(N-1)} \sum_{i=1}^N Y_u^2 - \frac{N-n}{(N-1)n} \bar{Y}^2 \\
&= \frac{N-n}{Nn} \cdot \frac{1}{N-1} \left[\sum_{i=1}^N Y_u^2 - N\bar{Y}^2 \right] \\
&= \frac{N-n}{Nn} S^2 = V_{\text{ran}}. \blacksquare
\end{aligned}$$

第二种解释是将总体看作是从一个无限的超总体 (super-population) 中随机抽取的一个样本量为 N 的样本. “平均”则是指按该超总体的概率分布取的期望值, 以 \mathcal{E} 表示之.

定理 8.4 若 $Y_u (u=1, 2, \dots, N)$ 是从某超总体中随机抽取的,

$$\mathcal{E}(Y_u) = \mu,$$

$$\mathcal{E}(Y_u - \mu)(Y_v - \mu) = \begin{cases} 0, & \text{若 } u \neq v; \\ \sigma_u^2, & \text{若 } u = v. \end{cases} \quad (8.9)$$

则

$$\mathcal{E}(V_{\text{sr}}) = \mathcal{E}(V_{\text{ran}}). \quad (8.10)$$

证明 对固定的有限总体

$$\begin{aligned}
V_{\text{ran}} &= \frac{N-n}{Nn} \cdot \frac{1}{N-1} \sum_{u=1}^N (Y_u - \bar{Y})^2 \\
&= \frac{N-n}{Nn} \cdot \frac{1}{N-1} \left[\sum_{u=1}^N (Y_u - \mu)^2 - N(\bar{Y} - \mu)^2 \right],
\end{aligned}$$

故

$$\begin{aligned}
\mathcal{E}(V_{\text{ran}}) &= \frac{N-n}{Nn(N-1)} \left[\sum_{u=1}^N \mathcal{E}(Y_u - \mu)^2 - N\mathcal{E}(\bar{Y} - \mu)^2 \right] \\
&= \frac{N-n}{Nn(N-1)} \left[\sum_{u=1}^N \sigma_u^2 - N\mathcal{E} \left\{ \frac{1}{N} \sum_{u=1}^N (Y_u - \mu)^2 \right\} \right] \\
&= \frac{N-n}{Nn(N-1)} \left[\sum_{u=1}^N \sigma_u^2 - \frac{1}{N} \sum_{u=1}^N \sigma_u^2 \right] \\
&= \frac{N-n}{N^2n} \sum_{u=1}^N \sigma_u^2.
\end{aligned}$$

另一方面

$$\begin{aligned}
\mathcal{E}(V_{\text{sr}}) &= \mathcal{E} \left[\frac{1}{k} \sum_{i=1}^k (\bar{Y}_i - \bar{Y})^2 \right] \\
&= \frac{1}{k} \mathcal{E} \left[\sum_{i=1}^k (\bar{Y}_i - \mu)^2 - k(\bar{Y} - \mu)^2 \right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{k} \left\{ \frac{1}{n^2} \mathcal{C} \sum_{i=1}^k \left[\sum_{j=1}^n (Y_{ij} - \mu) \right]^2 - \frac{k}{N^2} \sum_{u=1}^N \sigma_u^2 \right\} \\
&= \frac{1}{k} \left[\frac{1}{n^2} \sum_{u=1}^N \sigma_u^2 - \frac{k}{N^2} \sum_{u=1}^N \sigma_u^2 \right] \\
&= \frac{N-n}{N^2 n} \sum_{u=1}^N \sigma_u^2,
\end{aligned}$$

于是 $\mathcal{C}(V_{sy}) = \mathcal{C}(V_{ran})$. ■

8.3.2 线性趋势

若总体单元是按其指标值的大小顺序排列或按某个与其有线性相关的辅助变量 \mathcal{X} 的大小顺序排列, 此时 Y_u 与编号 u 也线性相关, 此种情况称为线性趋势排列, 也称 Y_u 按有关标识排列. 正如已在 8.2.4 段数值例子中所述的那样, 对按线性趋势排列的总体进行系统抽样能较大幅度地提高抽样精度, 原因是它增大了样本内方差.

本段我们将对线性趋势情形进行初步的定性说明, 所考虑的模型具有以下简单的形式:

$$Y_u = \alpha + \beta u \quad (8.11)$$

或

$$Y'_u = \frac{Y_u - \alpha}{\beta} = u. \quad (8.12)$$

以下仍用 Y_u 记 Y'_u .

定理 8.5 对于线性趋势模型 $Y_u = u (u=1, 2, \dots, N)$, 有

$$V_{st} \leq V_{sy} \leq V_{ran}. \quad (8.13)$$

证明 利用恒等式

$$\begin{aligned}
\sum_{u=1}^N u &= \frac{1}{2} N(N+1), \\
\sum_{u=1}^N u^2 &= \frac{1}{6} N(N+1)(2N+1)
\end{aligned}$$

可求得总体的方差为

$$S^2 = \frac{1}{12} N(N+1), \quad (8.14)$$

从而

$$V_{ran} = \frac{N-n}{Nn} S^2 = \frac{1}{12} (k-1)(N+1). \quad (8.15)$$

按分层随机抽样, 计算层内方差 S_u^2 的公式与计算 S^2 的公式形式完全相同, 只须用 k 代替 N 即可, 故

$$S_w^2 = \frac{1}{12} k(k+1). \quad (8.16)$$

因此根据比例分配分层随机抽样的方差公式, 有

$$V_{st} = \frac{N-n}{Nn} S_w^2 = \frac{1}{12n} (k^2-1). \quad (8.17)$$

对于系统抽样, 在此情形, 由于 k 个不同样本的样本均值 \bar{Y}_i , 依次都相差 1, 因而再次应用 (8.14) 式, 得

$$\frac{1}{k-1} \sum_{i=1}^k (\bar{Y}_i - \bar{Y})^2 = \frac{1}{12} k(k+1),$$

所以

$$V_{sy} = \frac{1}{k} \sum_{i=1}^k (\bar{Y}_i - \bar{Y})^2 = \frac{1}{12} (k^2-1). \quad (8.18)$$

比较 (8.15)、(8.17) 与 (8.18) 式即有

$$V_{st} \leq V_{sy} \leq V_{ran},$$

而且等号当且仅当 $n=1$ 时成立. ■

在实际问题中, 上述模型不可能严格成立, 但其结论在定性上还是适用的. 在下一节中将专门讨论线性趋势排列情形.

8.3.3 单元指标呈周期性变化的情形

在另外一些问题中, 总体单元指标 Y_u 呈周期性变化. 例如城市街道上的交通流量以 24 小时为周期相继出现高峰与低谷; 商店的销售额每周、每月以至每个季度都有周期性变化. 当 Y_u 随着 u 呈周期变化时, 系统抽样的效果, 即 V_{sy} 的大小与抽样间距 k 的选取有极大的关系.

假设 Y_u 的变化以 l 为周期, 则当 k 为 l 的整数倍时, 样本单元都取同一数值, 此时系统样本的代表性最差, 方差最大. 若取 $k=l-1$ 或 $l+1$, 则样本单元包含一个变化周期内许多有代表性的数值, 从而使方差大大减小, 因而精度较高. 因此, 当 Y_u 呈周期性变化时, 必须根据实际情况慎重地选择 k . 而这取决于抽样工作者对总体的了解和自身的经验.

8.3.4 单元指标呈自相关的情形

在有些情形, 特别是总体单元的排列是按空间位置或时间顺序排列时, 总体指标值 Y_i 之间存在一定的相关关系, 而且是正相关关系. 距离相近单元的相关较大, 距离较远的单元相关较小. 具体地说, 相关系数是

单元间距 u 的递减函数。下面我们假定总体单元 Y_1, Y_2, \dots, Y_N 是从一个满足以下条件的超总体中抽取的随机样本:

$$\mathcal{E}(Y_i) = \mu, \quad \mathcal{E}(Y_i - \mu)^2 = \sigma^2, \quad \mathcal{E}(Y_i - \mu)(Y_{i+u} - \mu) = \rho_u^2 \sigma^2. \quad (8.19)$$

其中 ρ_u 又满足 $\rho_u \geq \rho_v \geq 0$ (对 $u < v$). 这个模型即是一种自相关 (autocorrelated) 模型。对于自相关总体, 系统抽样有可能优于分层随机抽样。事实上, Cochran (1946) 证明了如下的结果 (仍设 $N = nk$):

定理 8.6 对于自相关总体 (8.19), 若又有

$$\delta_u^2 = \rho_{u+1} + \rho_{u-1} - 2\rho_u \geq 0 \quad (u = 2, 3, \dots, kn-2), \quad (8.20)$$

则

$$\mathcal{E}(V_{sy}) \leq \mathcal{E}(V_{st}) \leq \mathcal{E}(V_{ran}), \quad (8.21)$$

上式左端的等号仅在 $\delta_u^2 = 0$ ($u = 2, 3, \dots, kn-2$) 时才成立。

证明 我们仅须证明 $\mathcal{E}(V_{sy}) \leq \mathcal{E}(V_{st})$, 其中

$$\mathcal{E}(V_{sy}) = \mathcal{E}E(\bar{y}_{sy} - \bar{Y})^2.$$

对于分层样本, 每层中的样本单元有 k 个可能的 (相对) 位置, 从而任意两层的入样单元在层中的位置共有 k^2 种可能的组合。可能的距离为 $1, 2, \dots, k-1, k, k+1, \dots, 2k-2, 2k-1$; 而每种距离可能的位置组合数分别为 $1, 2, \dots, k-1, k, k-1, \dots, 2, 1$ 。因而 V_{sy} 的期望值可以写成:

$$\mathcal{E}(V_{st}) = \frac{\sigma^2}{2k^2} \left[\sum_{u=1}^{k-1} u(2 + \rho_u + \rho_{2k-u}) + k(1 + \rho_k) \right],$$

$$\text{而} \quad \mathcal{E}(V_{sy}) = \frac{\sigma^2}{2k^2} \left[\sum_{u=1}^{k-1} u(2 + 2\rho_k) + k(1 + \rho_k) \right].$$

由于 (8.20) 式, 故有

$$\mathcal{E}(V_{st}) - \mathcal{E}(V_{sy}) = \frac{\sigma^2}{2k^2} \left[\sum_{u=1}^{k-1} u(\rho_u + \rho_{2k-u} - 2\rho_k) \right] \geq 0.$$

而等号仅当对每个 u , $\rho_u + \rho_{2k-u} - 2\rho_k = \delta_k^2 = 0$ 时才成立。■

Quenouilli (1949) 证明了 (8.19) 中的前两个假定可放宽至

$$\mathcal{E}(Y_i) = \mu_i, \quad \mathcal{E}(Y_i - \mu_i)^2 = \sigma_i^2.$$

不少作者提出了在实际问题中可用的 ρ_u 的形式, 例如将

$$\rho_u = \tanh(u^{-\frac{3}{8}})$$

用来描述相距为 u 的两个气象台的降雨量的相关; $\rho_u = e^{-\lambda u}$ 用于农林土地调查, $\rho_u = (t-u)l$ 用于某些类型的时间序列等。

§ 8.4 具有线性趋势的总体的估计量与抽样方法的改进

上节已指出当总体单元的排列具有线性趋势时, 等距的系统抽样具有较高的精度. 由于这种排列的系统样本内方差增大, 故估计量的方差小于简单随机抽样的方差. 但是与按大小为层的分层随机抽样比较, 它的方差仍稍高. 这是因为系统抽样仅是一种样本单元的位置完全固定的分层抽样. 若初始单元的值在层内偏小或偏大, 则整个样本, 从而估计量的值也偏小或偏大, 这就增大了估计量的方差. 而在分层随机抽样中, 由于样本单元在层内的位置是随机的, 因此估计量的方差较小. 受此启发, 我们可以对系统抽样的估计量以至它的抽样方法作适当的改进, 以进一步提高其精度. 实际表明使用这些方法, 有可能使系统抽样达到比分层随机抽样更高的精度. 事实上, 以下介绍的大多数方法对于遵从严格而简单的线性趋势模型(8.11)的总体可完全消除其线性趋势的影响, 即使方差减少到 0 的理想情形.

8.4.1 首尾校正法

Yates(1948) 首先对 $N = nk$ 的情形提出在计算估计量即样本均值时, 采用加权平均, 对首尾两个样本单元赋以与其他单元不同的权. 设初始单元的编号为 i , 则样本中所有中间单元的权 w_j 仍为 $\frac{1}{n}$ ($j = 2, \dots, n-1$), 而首尾两个单元的权分别取为:

$$w_1 = \frac{1}{n} + \frac{2i - k - 1}{2(n-1)k}, \quad w_n = \frac{1}{n} - \frac{2i - k - 1}{2(n-1)k}. \quad (8.22)$$

经过上述修正后, 在模型(8.11)时, 对任何 i , 都有

$$\bar{y}_s = \sum_{j=1}^n w_j Y_{ij} = \bar{Y}. \quad (8.23)$$

对于 $N \neq nk$ 的情形, Bellhouse 与 Rao(1975) 也提出了类似的修正. 假定按 Lahiri 的圆形系统抽样法抽取样本, 按总体单元原有排列顺序确定首尾两个样本单元(不是指抽样过程中的顺序). 若抽样时初始单元的编号 i 较小, 满足 $i + (n-1)k \leq N$, 此时所有 n 个样本单元都不超过单元 N , 则对首尾两个样本单元赋予以下的权:

$$\begin{aligned}
 w_1 &= \frac{1}{n} + \frac{2i + (n-1)k - (N+1)}{2(n-1)k}, \\
 w_n &= \frac{1}{n} - \frac{2i + (n-1)k - (N+1)}{2(n-1)k}.
 \end{aligned} \tag{8.24}$$

若 $i + (n-1)k > N$, 则必有取到的样本单元越过单元 N , 设越过单元 N 的样本单元数为 n_2 , 则相应的权取为:

$$\begin{aligned}
 w_1 &= \frac{1}{n} + \frac{2i + (n-1)k - (N+1) - 2n_2 N/n}{2(N-k)}, \\
 w_n &= \frac{1}{n} - \frac{2i + (n-1)k - (N+1) - 2n_2 N/n}{2(N-k)}.
 \end{aligned} \tag{8.25}$$

其他所有中间单元的权仍为 $\frac{1}{n}$.

8.4.2 中位样本法

Madow(1953)提出为消除系统抽样中初始单元位置的影响, 固定取层内处于中间位置的样本点, 也即令

$$i = \begin{cases} (k+1)/2, & \text{若 } k \text{ 为奇数;} \\ \frac{k}{2} \text{ 或 } \frac{k}{2} + 1, & \text{若 } k \text{ 为偶数.} \end{cases}$$

就一次调查而言, 中位样本法的效果较好. 但缺点是: 按照这种方法, 样本不再是随机的了. 总体单元排列顺序一旦确定, 样本也就确定了. 因此对同样问题进行多次定时调查时, 这种抽样会带来不利的影响.

8.4.3 对称(平衡)系统抽样法

另一种改进方法的思想是初始单元不是一个, 而是两个, 位置对称, 数值大小相抵, 从而减小估计量的方差. 这种方法通称为对称系统抽样或平衡系统抽样(balanced systematic sampling). 具体方法有两种. 我们首先都假定 $N = nk$, 且 n 为偶数. 第一种方法是由 Sethi(1965)最先提出, 后经 Murthy(1967)总结的方法, 将总体分为 $n/2$ 层, 每层包含 $2k$ 个单元. 在每层中随机确定与两端等距的两个单元作为样本单元, 每层中的样本单元位置一致. 具体地说, 当起始随机数为 $i (1 \leq i \leq k)$, $\frac{n}{2}$ 对样本单元的入样号码为:

$$[i + 2jk, 2(j+1)k - i + 1] \quad \left(j = 0, 1, 2, \dots, \frac{n}{2} - 1 \right). \tag{8.26}$$

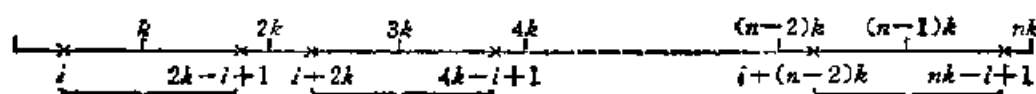


图 8.2 层内对称系统抽样

这种抽样可用图 8.2 表示。由于样本单元在层内的位置是对称的，因此，我们称这种对称系统抽样为层内对称系统抽样。

Singh 等(1968)对上述方法作了修正，提出另一种对称系统抽样法，仍设 n 为偶数，当确定一个 $[1, k]$ 之间的随机整数 i 后， $n/2$ 对样本单元由以下确定：

$$[i + jk, N - i - jk + 1] \quad (j = 0, 1, 2, \dots, \frac{n}{2} - 1). \quad (8.27)$$

每对样本单元在总体中的位置都是对称的，因此，我们将这种方法称为总体对称系统抽样，如图 8.3 所示。

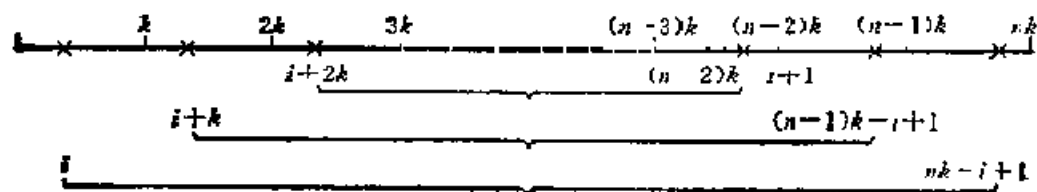


图 8.3 总体对称系统抽样

若 n 为奇数，则样本单元总有一个是不成对的。这个单元通常取为总体中的中间位置。即在层内对称系统抽样中，从总体排列的两端起分层，最后留下中间的“半层”，在这半层(包含 k 个单元)中随机地抽取一个单元，或干脆就取中间位置的单元作为样本单元。在总体对称系统抽样的情形也作同样处理，在剩下的 k 个总体单元中随机地或抽取中间位置的单元作为样本单元。

8.4.4 回归估计量的应用

从理论上说，Yates 的首尾修正法及 Sethi 与 Singh 等提出的两种对称系统抽样对于完全线性趋势又不存在随机误差的模型(8.11)，都能完全消除其线性影响(假定 $N = nk$ ，且 n 为偶数的情形，若不然，则稍有误差)。但在实际应用中最常见的线性趋势模型是带随机误差的。因此一般的线性趋势模型是将总体单元 $Y_i (i = 1, 2, \dots, N)$ 看成是从以下超总体中抽取的随机样本：

$$Y_i = \mu_i + \varepsilon_i, \quad \mu_i = \alpha + \beta X_i + \varepsilon_i, \quad (8.28)$$

其中

$$E(\varepsilon_i) = 0,$$

$$\mathcal{C}(s_i s_j) = \begin{cases} \sigma_i^2, & \text{若 } i = j; \\ 0, & \text{若 } i \neq j. \end{cases} \quad (8.29)$$

式中 X_i 是某个辅助变量, 更一般的, 我们可以定义二次趋势模型等, 对于这类模型我们可以应用第4章讨论过的回归估计, 对于所得的系统样本, 不用样本平均数的简单估计, 而用(例如)线性回归估计:

$$\bar{y}_{lr} = \bar{y}_{sv} + \beta(\bar{X} - \bar{x}), \quad (8.30)$$

更一般的, 如果 $f(\cdot)$ 是一个已知函数, 则可定义以下估计:

$$\hat{\bar{Y}} = \bar{y}_{sv} + \frac{1}{N} \sum_{i=1}^N f(X_i) - \frac{1}{n} \sum_{i=1}^n f(x_i). \quad (8.31)$$

可以证明对于线性趋势模型(8.28), 有

$$\mathcal{C}E(\bar{y}_{lr} - \bar{Y})^2 = \frac{N-n}{N^2 n} \sum_{i=1}^N \sigma_i^2 \leq \mathcal{C}E(\bar{y}_{sv} - \bar{Y})^2. \quad (8.32)$$

因此, 在期望均方误差(对于无偏估计, 即是期望方差)的标准下, 系统样本的回归估计比通常的简单估计的精度要高.

8.4.5 数值例子——部门职工总人数的估计

为估计某部门当年的职工总人数 Y , 将该部门各单位按上一年职工统计人数 X_i 从小到大的顺序排列, 按等距抽样的首尾校正法, 两种对称系统抽样法以及对后两种方法所得样本的线性回归估计比较其方差. 为简明起见, 我们仅对一个子总体($N=32$), 按 $n=8$, $b=4$ 作模拟抽样, 原始数据如表 8.6 所示.

表 8.6 某部门各单位上一年职工人数 X_i 与当年职工人数 Y_i

单位编号 i	X_i	Y_i	单位编号 i	X_i	Y_i
1	45	48	17	199	206
2	48	50	18	210	218
3	59	66	19	222	243
4	63	74	20	245	248
5	76	78	21	268	263
6	90	107	22	291	301
7	97	115	23	324	326
8	102	123	24	350	358
9	114	111	25	382	395
10	118	130	26	394	402
11	127	135	27	416	429
12	140	142	28	423	435
13	144	148	29	458	467
14	162	152	30	473	499
15	174	183	31	510	532
16	196	200	32	562	579

为方便起见,我们将估计的总体参数取为均值 \bar{Y} . 根据表 8.6 中的数据, $\bar{Y} = 242.59375$, $S_y^2 = 24011.217$, 又 $\bar{X} = 233.96875$, Y_i 对 X_i 的(总体)回归系数 $\beta = 1.021068$.

对每种方法,可能的系统样本都为 4 个. 我们计算其均值及均方误差的平方根. 后者计算公式为

$$\sqrt{\text{MSE}(\hat{\theta})} = \sqrt{\frac{1}{4} \sum_{i=1}^4 (\hat{\theta}_i - \bar{Y})^2}.$$

一、一般系统抽样(等距抽样)

4 个可能样本所包含的单元号以及估计量列于表 8.7 中, 其中估计量又分未经校正的通常简单估计量 \bar{y}_{sy} 及已经 Yates 的首尾校正后的估计量 \bar{y}'_{sy} . 从表中可见, \bar{y}_{sy} 是无偏的, 而 \bar{y}'_{sy} 是有偏的, 但后者的 $\sqrt{\text{MSE}}$ 要比前者小得多.

表 8.7 表 8.6 总体等距抽样的全部可能样本及其估计量 $\hat{\theta}$

样本序号	样本所包含的单元顺序号	y_{sy}	\bar{y}'_{sy}
1	1, 5, 9, 13, 17, 21, 25, 29	214.500	236.94643
2	2, 6, 10, 14, 18, 22, 26, 30	232.375	240.39286
3	3, 7, 11, 15, 19, 23, 27, 31	253.625	245.30357
4	4, 8, 12, 16, 20, 24, 28, 32	269.875	242.82143
	$E(\hat{\theta})$	242.59375	241.36607
	$\sqrt{\text{MSE}(\hat{\theta})}$	20.97401	3.32156

二、层内对称系统抽样

4 个可能样本所包含的单元号以及简单估计量 \bar{y}_{sy} 与线性回归估计量 \bar{y}_{lr} 列于表 8.8 中. 两种估计量都是无偏的(回归估计量中所用的 β 是总体值而非样本估计值), 但显然回归估计量的标准差比简单估计的标准差小得多. 虽然按这种抽样方法的简单估计本身又比一般系统抽样的简单估计的精度要高.

三、总体对称(修正)系统抽样

4 个可能样本所包含的单元号以及简单估计量 y_{sy} 与线性回归估计量 \bar{y}_{lr} 列于表 8.9 中, 从表中的数据可见, 这种抽样方法的效果与层内对称系统抽样相差不多.

表 8 8 表 8.6 总体按层内对称系统抽样的全部可能样本的简单估计与回归估计

样本序号	样本所包含的单元顺序号	\bar{y}_{sy}	\bar{y}_{lr}
1	1, 8, 9, 16, 17, 24, 25, 32	252.500	242.513
2	2, 7, 10, 15, 18, 23, 26, 31	244.500	244.085
3	3, 6, 11, 14, 19, 22, 27, 30	241.500	245.552
4	4, 5, 12, 13, 20, 21, 28, 29	231.875	238.225
	$E(\hat{\theta})$	242.59375	242.59375
	$\sqrt{V(\hat{\theta})}$	7.37996	2.74170

表 8 9 表 8.6 的总体按总体对称系统抽样的全部可能样本的简单估计与回归估计

样本序号	样本所包含的单元顺序号	\bar{y}_{sy}	\bar{y}_{lr}
1	1, 5, 9, 13, 20, 24, 28, 32	250.625	239.489
2	2, 6, 10, 14, 19, 23, 27, 31	246.125	243.796
3	3, 7, 11, 15, 18, 22, 26, 30	239.875	245.842
4	4, 8, 12, 16, 17, 21, 25, 29	233.750	241.248
	$E(\hat{\theta})$	242.59375	242.59375
	$\sqrt{V(\hat{\theta})}$	6.37523	2.42111

从本例中我们可以看到,对于按辅助变量大小顺序排列的总体(可用线性趋势或二次趋势模型近似),用对称系统抽样的效果显然优于一般的等距抽样.而两种对称系统抽样与经 Yates 首尾校正法之间的优劣不可一概而论,它主要取决于总体结构.特别是 Y_i 与 X_i 的相关程度以及 X_i 的内部结构,而回归估计量的效果正如预期的那样,一般更好一些,不过它是以付出更多的计算量为代价的.

§ 8.5 不等概率系统抽样

8.5.1 概述及实施方法

不等概率系统抽样是使用最为广泛的不放回不等概率抽样方法之

一. 它之所以受欢迎, 主要是因为它结合了系统抽样方便易行与不等概率抽样的高效率的共同特点. 作为一种不放回的不等概率抽样, 它很容易地成为一种 π PS 抽样, 其方法也适用于任意样本量 n 的情形. 这与许多实用的 π PS 抽样仅适用于 $n=2$ 的情况完全不同. 因此不等概率系统抽样的总体效率较高. 不过与其他一些系统抽样一样, 它的方差估计是较为困难的.

对总体的 N 个单元的某种确定的排列顺序, 若 $\{\pi_i, i=1, 2, \dots, N\}$ 是一组包含概率, $\sum_{i=1}^N \pi_i = n$, 不等概率系统抽样的一般方法是先在 $[0, 1]$ 范围内随机地抽取一个实数 r , 则满足下列条件的总体中的第 i_0, i_1, \dots, i_{n-1} 个单元入样:

$$\sum_{j=1}^{i-1} \pi_j < r + k, \quad \sum_{j=1}^{i_2} \pi_j \geq r + k \quad (k=0, 1, \dots, n-1). \quad (8.33)$$

在应用中最为常用的是 π PS 系统抽样, 也即入样概率与单元大小 M_i 成比例的系统抽样. 令 $M_0 = \sum_{i=1}^N M_i$, 则 $\pi_i = \frac{nM_i}{M_0}$. 具体进行抽样则与一般的系统抽样类似, 也用通常 PPS 抽样中的代码法. 对第一个单元赋以 $1 \sim M_1$ 共 M_1 个代码; 对第二个单元赋以 $M_1+1 \sim M_1+M_2$ 共 M_2 个代码, \dots , 对第 i 个单元赋以 $\sum_{j=1}^{i-1} M_j + 1 \sim \sum_{j=1}^i M_j$ 共 M_i 个代码 \dots . 令 k 为最接近于 M_0/n 的整数 (不失一般性, 我们设 M_i 皆为整数) 则从 $1 \sim k$ 范围内随机地产生一个整数 r , 则代码 $r, r+k, \dots, r+(n-1)k$ 所在的单元即为入样单元.

例 8.8 设总体由表 8.10 中 $N=8$ 个单元组成, $M_0=45$, 若 $n=3$, $k=15$, 又 $1 \sim 15$ 范围内产生的随机数 $r=5$, 则代码为 5, 20, 35 的三个单

表 8.10 π PS 系统抽样示例

i	M_i	$\sum_{j=1}^i M_j$
1	5	5
2	10	15
3	2	17
4	8	25
5	9	34
6	3	37
7	1	38
8	7	45
Σ	$M=45$	

元,即第 1, 4, 6 三个单元入样.

注意:上述抽样方法还不能保证抽样是不放回的.事实上,对于那些特别大的单元,也即 $M_i > k$ 的单元,对某些 r ,有可能被重复抽到.为避免这种情形发生,最好的方法是将这些单元事先抽出来,对每个都进行调查.然后在其他单元组成的子总体中再进行抽样.这样做不仅保证了方法是不放回的,而且效率更高.

8.5.2 估计量

对于不等概率系统抽样,总体总和 Y 的估计仍可用通常不放回的不等概率抽样中的 Horvitz-Thompson 估计量:

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i}. \quad (8.34)$$

根据定理 5.2, \hat{Y}_{HT} 是 Y 的无偏估计,其方差为:

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \frac{1-\pi_i}{\pi_i} Y_i^2 + 2 \sum_{i=1}^N \sum_{j=1}^N \frac{\pi_{ij}}{\pi_i \pi_j} Y_i Y_j. \quad (8.35)$$

当 n 固定时,又可以表示为:

$$V(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{j=1}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2. \quad (8.36)$$

但上述方差并不一定能根据定理 5.3 用样本进行估计,其原因是在不等概率系统抽样中,并不总能保证对所有的 $\pi_{ij} > 0$. 事实上,通常都有不少 $\pi_{ij} = 0$. 例如在例 8.3 中, $\pi_{12} = \pi_{23} = \pi_{34} = \pi_{56} = \pi_{57} = \pi_{67} = \pi_{68} = 0$. 因此方差估计必须用别的方法. 我们将在 § 8.6.2 中讨论这个问题.

§ 8.6 系统抽样中的方差估计

与前几章讨论过的其他基本抽样方法不同的是:系统抽样估计量的方差估计没有理想的和精确的方法. 本节介绍的许多估计方法在某种程度上都是近似的,在实际应用时要区别情况,对不同的总体模型选择较为合适的估计量.

8.6.1 等概率系统抽样的情形

对于等概率系统抽样,即一般意义的等距抽样,有八种可用的方差估计. 我们从直观上解释这些估计的构造思想,并进行比较,指出各自的使用场合. 与以前一样,我们仍然假定 $N = nk$, 在必要时进一步假定 n 为偶

数. 另外, 为表达方便, 我们再次将 N 个总体单元按行(群)、列(层)排列, 记为 Y_{ij} , 且抽取的初始单元编号为 i . 我们讨论的是对 \bar{Y} 的估计量

$\bar{y}_{sy} = \frac{1}{n} \sum_{j=1}^n Y_{ij}$ 的方差 $V(\bar{y}_{sy})$ 的估计.

一、八种方差估计

若将系统样本视为简单随机样本, 则 $V(\bar{y}_{sy})$ 的估计可用

$$v_1(i) = \frac{1-f}{n} s^2, \quad (8.37)$$

其中 $f = \frac{n}{N}$, $s^2 = \frac{1}{n-1} \sum_{j=1}^n (Y_{ij} - \bar{y}_{sy})^2$.

若从第二个样本单元起都与前一个样本单元组成一对, 共 $n-1$ 对. 每对单元的(样本)方差可表示为 $\frac{1}{2} (Y_{ij} - Y_{i,j-1})^2$, 因此 $V(\bar{y}_{sy})$ 的方差估计可表示为:

$$v_2 = \frac{1-f}{n} \sum_{j=2}^n a_{ij}^2 / [2(n-1)], \quad (8.38)$$

其中

$$a_{ij} = \Delta Y_{ij} = Y_{ij} - Y_{i,j-1}. \quad (8.39)$$

如果仅考虑相邻不重迭的两个样本单元对, 共 $n/2$ 对, 可得

$$v_3 = \frac{1-f}{n} \sum_{j=1}^{n/2} a_{i,2j}^2 / n. \quad (8.40)$$

v_2 、 v_3 只是考虑了样本观测值的一阶差分, 可以考虑用更高阶的差分, 于是有以下三种方差估计:

$$v_4(i) = \frac{1-f}{n} \sum_{j=3}^n b_{ij}^2 / [6(n-2)], \quad (8.41)$$

$$v_5(i) = \frac{1-f}{n} \sum_{j=5}^n c_{ij}^2 / [3.5(n-4)], \quad (8.42)$$

$$v_6(i) = \frac{1-f}{n} \sum_{j=9}^n d_{ij}^2 / [7.5(n-8)], \quad (8.43)$$

其中

$$b_{ij} = \Delta^2 Y_{ij} = Y_{ij} - 2Y_{i,j-1} + Y_{i,j-2}, \quad (8.44)$$

$$\begin{aligned} c_{ij} &= \frac{1}{2} \Delta^4 Y_{ij} + \Delta^2 Y_{i,j-1} \\ &= \frac{1}{2} Y_{ij} - Y_{i,j-1} + Y_{i,j-2} - Y_{i,j-3} + \frac{1}{2} Y_{i,j-4}, \end{aligned} \quad (8.45)$$

$$\begin{aligned}
 d_{ij} &= \frac{1}{2} \Delta^8 Y_{i,j} + 3\Delta^6 Y_{i,j-1} + 5\Delta^4 Y_{i,j-2} + 2\Delta^2 Y_{i,j-3} \\
 &= \frac{1}{2} Y_{i,j} - Y_{i,j-1} + Y_{i,j-2} - Y_{i,j-3} + \cdots - Y_{i,j-7} + \frac{1}{2} Y_{i,j-8}.
 \end{aligned} \quad (8.46)$$

注意 v_2, v_4, v_5 与 v_6 分母中的系数 2, 6, 3.5, 7.5 分别等于 $a_{ij}, b_{ij}, c_{ij}, d_{ij}$ 对 Y_{ij} 展开式中各项系数的平方和.

若将样本随机地分成 m 个子样本, 每个包含 n/m 个单元 (假定 n/m 为整数), 令 \bar{y}_α 为第 α 个子样本的平均数, 则

$$\bar{y}_{sy} = \frac{1}{m} \sum_{\alpha=1}^m \bar{y}_\alpha,$$

于是 $V(\bar{y}_{sy})$ 也可用下式估计:

$$v_7(i) = \frac{1-f}{m(m-1)} \sum_{\alpha=1}^m (\bar{y}_\alpha - \bar{y}_{sy})^2. \quad (8.47)$$

这一方法称为随机分组法, 在下一章中将详细地讨论一般情形下的这一类方差估计.

最后一个估计量是用总体中相距为 k 的两个单元的相关系数 ρ_k 来表示的. 事实上, 可以用此构造一类估计量. v_8 是由 Cochran (1946) 提出的:

$$v_8(i) = \begin{cases} \frac{1-f}{n} s^2 \left[1 + \frac{2}{\ln \hat{\rho}_k} + \frac{2}{\hat{\rho}_k^2 - 1} \right], & \text{若 } \hat{\rho}_k > 0; \\ -\frac{1-f}{n} s^2, & \text{若 } \hat{\rho}_k \leq 0. \end{cases} \quad (8.48)$$

其中 $\hat{\rho}_k$ 是 ρ_k 的估计.

二、不同估计量的比较

为研究上述八种方差估计量在不同模型下的性质, 我们考虑一些典型的超总体模型.

假定所考察的总体是由以下超总体模型随机产生的:

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}. \quad (8.49)$$

其中 μ_{ij} 是 i, j 的已知 (常数) 函数, ε_{ij} 是随机分量. 对 μ_{ij} 的不同函数形式就有不同的超总体模型. 对 ε_{ij} 我们假定

$$\mathcal{E}(\varepsilon_{ij}) = 0, \quad \mathcal{E}(\varepsilon_{ij}^2) = \sigma^2. \quad (8.50)$$

$V(y_{sy})$ 的某个估计 v_α (在超总体意义下) 的期望偏倚和相对期望偏倚分别定义为:

$$\mathcal{B}(v_\alpha) = \mathcal{E}E(v_\alpha) - \mathcal{E}V(y_{sy}), \quad (8.51)$$

$$\mathcal{R}(v_a) = \mathcal{B}(v_a) / \mathcal{E}V(\bar{y}_{sy}). \quad (8.52)$$

1. 随机模型(random model) 若对所有的 i, j, μ_{ij} 皆为常数, 即

$$\mu_{ij} = \mu, \quad (8.53)$$

且 ε_{ij} 是独立同分布的随机变量, 则模型称为随机模型. 对这个模型, y_{sy} 的期望方差是

$$\mathcal{E}V(\bar{y}_{sy}) = \frac{1-f}{n} \sigma^2. \quad (8.54)$$

可以证明 $V(\bar{y}_{sy})$ 的前七个估计的期望偏倚皆为零, $\mathcal{B}(v_8)$ 也接近于零. 因此对于随机模型, 所有八种估计的效果都不错, 但鉴于 v_1 最简单, 故它是最佳选择.

2. 线性趋势模型(linear trend model) 这个模型的 μ_{ij} 有表达式:

$$\mu_{ij} = \beta_0 + \beta_1[i + (j-1)k]. \quad (8.55)$$

其中 β_0, β_1 是未知常数, 且 ε_{ij} 也是独立同分布的随机变量. 此时

$$\mathcal{E}V(\bar{y}_{sy}) = \beta_1^2(k^2 - 1)/12 + (1-f)\sigma^2/n. \quad (8.56)$$

各个方差估计的期望值分别为

$$\mathcal{E}E(v_1) = (1-f)[\beta_1^2 k^2(n+1)/12 + \sigma^2/n], \quad (8.57)$$

$$\mathcal{E}E(v_2) = \mathcal{E}E(v_3) = (1-f)[\beta_1^2 k^2/(2n) + \sigma^2/n], \quad (8.58)$$

$$\mathcal{E}E(v_4) = \mathcal{E}E(v_5) = \mathcal{E}E(v_6) = (1-f)\sigma^2/n, \quad (8.59)$$

$$\mathcal{E}E(v_7) = (1-f)[\beta_1^2 k^2(m+1)/12 + \sigma^2/n], \quad (8.60)$$

$$\mathcal{E}E(v_8) \approx (1-f)[\gamma(0)/n] \cdot \left[1 + \frac{2}{\ln\{\gamma(1)/\gamma(0)\}} + \frac{2}{\gamma(0)/\gamma(1) - 1} \right]. \quad (8.61)$$

其中

$$\gamma(1) = \beta_1^2 k^2(n-3)(n+1)/12 + \sigma^2/n,$$

$$\gamma(0) = \beta_1^2 k^2 n(n+1)/12 + \sigma^2. \quad (8.62)$$

从上述公式可看到, 对于较大的 k , 且 β_1 不十分接近于零时, 有

$$\mathcal{R}(v_1) \approx n,$$

$$\mathcal{R}(v_2) = \mathcal{R}(v_3) \approx -(n-6)/n,$$

$$\mathcal{R}(v_4) = \mathcal{R}(v_5) = \mathcal{R}(v_6) \approx -1,$$

$$\mathcal{R}(v_7) \approx k.$$

故从相对期望偏倚的观点看, v_2 与 v_3 最好, 其次是 v_4, v_5 与 v_6 . 模拟结果表明 v_8 有时也相当不错.

3. 分层效应模型(stratification effects model) 这个模型中的 μ_{ij} 满足

$$\mu_{ij} = \mu_j. \quad (8.63)$$

s_{ij} 仍是独立同分布的随机变量, 此时

$$\mathcal{E}V(\bar{y}_{sy}) = (1-f)\sigma^2/n. \quad (8.64)$$

各个方差估计的期望值为(其中 $\bar{\mu} = \sum_{j=1}^n \mu_j/n$):

$$\mathcal{E}E(v_1) = (1-f) \left\{ \sum_j^n (\mu_j - \bar{\mu})^2 / [n(n-1)] + \sigma^2/n \right\}, \quad (8.65)$$

$$\mathcal{E}E(v_2) = (1-f) \left\{ \sum_j^{n-1} (\mu_j - \mu_{j+1})^2 / [2n(n-1)] + \sigma^2/n \right\}, \quad (8.66)$$

$$\mathcal{E}E(v_3) = (1-f) \left\{ \sum_j^{n/2} (\mu_{2j-1} - \mu_{2j})^2 / n^2 + \sigma^2/n \right\}, \quad (8.67)$$

$$\mathcal{E}E(v_4) = (1-f) \left\{ \sum_j^{n-2} (\mu_j - 2\mu_{j+1} + \mu_{j+2})^2 / [6n(n-2)] + \sigma^2/n \right\}, \quad (8.68)$$

$$\mathcal{E}E(v_5) = (1-f) \left\{ \sum_j^{n-4} (\mu_j/2 - \mu_{j+1} + \mu_{j+2} - \mu_{j+3} + \mu_{j+4}/2)^2 / [3.5n(n-4)] + \sigma^2/n \right\}, \quad (8.69)$$

$$\mathcal{E}E(v_6) = (1-f) \left\{ \sum_j^{n-8} (\mu_j/2 - \mu_{j+1} + \cdots - \mu_{j+7} + \mu_{j+8}/2)^2 / [7.5n(n-8)] + \sigma^2/n \right\}, \quad (8.70)$$

$$\mathcal{E}E(v_7) = (1-f) \left[m^{-1}(m-1)^{-1} \sum_{\alpha}^m (\bar{\mu}_{\alpha} - \bar{\mu})^2 + \sigma^2/n \right]. \quad (8.71)$$

其中 $\bar{\mu}_{\alpha}$ 是 μ_j 的第 α 个子样本的均值.

$$\mathcal{E}E(v_8) \approx (1-f)n^{-1}(\kappa(0) + \sigma^2) \left[1 + \frac{2}{\ln \frac{\kappa(1)}{\kappa(0) + \sigma^2}} + \frac{2}{\frac{\kappa(0) + \sigma^2}{\kappa(1)} - 1} \right], \quad (8.72)$$

其中

$$\kappa(0) = (n-1)^{-1} \sum_j^n (\mu_j - \bar{\mu})^2,$$

$$\kappa(1) = (n-1)^{-1} \sum_j^{n-1} (\mu_j - \bar{\mu})(\mu_{j+1} - \bar{\mu}).$$

从上述列出的结果看, 当 μ_j 相差不大时, 前七个估计量的相对偏倚都较小, 且大致相等; 当 μ_j 相差较大时, 不同估计量的效果有较大的差别: v_1 与 v_8 常常较大, 而 v_5 与 v_6 一般较好.

4. 自相关模型 (autocorrelated model) 与其他模型一个最大的区别是: 对于自相关模型, s_{ij} 不是相互独立的, 而是相关的, 我们考虑一个较简单的情况, 即一阶自相关模型:

$$Y_u - \mu = \rho(Y_{u-1} - \mu) + \varepsilon_u \quad (8.73)$$

其中 $u=1, 2, \dots, N$; ρ 是一阶自回归系数 ($0 < \rho < 1$). 可以证明

$$\mathcal{E}V(\bar{y}_{sy}) = \frac{1-f}{n} \sigma^2 \left(1 - \frac{2}{k-1} \cdot \frac{\rho}{1-\rho} + \frac{2k}{k-1} \cdot \frac{\rho^k}{1-\rho^k} \right) + O(n^{-2}). \quad (8.74)$$

各个估计量的期望值分别为:

$$\begin{aligned} \mathcal{E}E(v_1) &= (1-f)(\sigma^2/n) \left\{ 1 - \frac{2}{n-1} \cdot \frac{\rho^k - \rho^N}{(1-\rho^k)} + \frac{2}{n(n-1)} \right. \\ &\quad \times \left[\frac{\rho^k - \rho^N}{(1-\rho^k)^2} - \frac{(n-1)\rho^N}{1-\rho^k} \right] \Big\} \\ &= \frac{1-f}{n} \sigma^2 + O(n^{-2}), \end{aligned} \quad (8.75)$$

$$\mathcal{E}E(v_2) = \mathcal{E}E(v_3) = (1-f)(\sigma^2/n)(1-\rho^k), \quad (8.76)$$

$$\mathcal{E}E(v_4) = (1-f)(\sigma^2/n) [1 - 4\rho^k/3 + \rho^{2k}/3], \quad (8.77)$$

$$\mathcal{E}E(v_5) = (1-f)(\sigma^2/n) [1 - 12\rho^k/7 + 8\rho^{2k}/7 - 4\rho^{3k}/7 + \rho^{4k}/7], \quad (8.78)$$

$$\begin{aligned} \mathcal{E}E(v_6) &= (1-f)(\sigma^2/n) [1 - 28\rho^k/15 + 24\rho^{2k}/15 - 20\rho^{3k}/15 \\ &\quad + 16\rho^{4k}/15 - 12\rho^{5k}/15 + 8\rho^{6k}/15 - 4\rho^{7k}/15 + \rho^{8k}/15], \end{aligned} \quad (8.79)$$

$$\begin{aligned} \mathcal{E}E(v_7) &= (1-f)(\sigma^2/n) \left\{ 1 + [2/(m-1)] [m(\rho^{mk} - \rho^N)/(1-\rho^{mk}) \right. \\ &\quad - (\rho^k - \rho^N)/(1-\rho^k)] - [2/(m-1)] \\ &\quad \times \left[\frac{m^2}{n} ((\rho^{mk} - \rho^N)/(1-\rho^{mk})^2 - (n/m-1)\rho^N/(1-\rho^{mk})) \right. \\ &\quad \left. \left. - n^{-1}((\rho^k - \rho^N)/(1-\rho^k)^2 - (n-1)\rho^N/(1-\rho^k)) \right] \right\} \\ &= \frac{1-f}{n} \sigma^2 \left[1 + \frac{2}{m-1} \left(\frac{m\rho^{km}}{1-\rho^{km}} - \frac{\rho^k}{1-\rho^k} \right) \right] + O(n^{-2}), \end{aligned} \quad (8.80)$$

$$\mathcal{E}E(v_8) = (1-f)(\sigma^2/n) [1 + 2/\ln(\rho^k) + 2\rho^k/(1-\rho^k)] + O(n^{-2}). \quad (8.81)$$

由此可知, 若 $\rho \approx 0$, 则八个估计量的偏倚都较小, 如果 k 较大, 则不论 ρ 取什么值 (除非 $\rho \approx 1$), 每个估计量的偏倚也不会很大. v_4 、 v_5 与 v_6 的效果不错. 而由于 $2/\ln(\rho^k)$ 是 $-2\rho/k(1-\rho)$ 的一个很好的近似, 因此 $\mathcal{E}E(v_8)$ 与 $\mathcal{E}V(\bar{y}_{sy})$ 几乎相等, 故 v_8 对于自相关总体是一个相当好的估计量.

5. 周期总体模型(periodic population model) 对于周期变化模型, μ_{ij} 是 $i + (j-1)k$ 的周期函数, 而 s_{ij} 是相互独立的随机变量. 一个简单的模型是

$$\mu_{ij} = \beta_0 \sin\{\beta_1[i + (j-1)k]\}.$$

正如我们在 § 8.3 中已经注意到的, 对周期变化的总体, 采用系统抽样要特别小心. 例如当抽样间距 k 是周期 $2\pi/\beta_1$ 的倍数时, \bar{y}_{sy} 的实际方差很大, 而所有八个估计量都很小. 反之, 当 k 为半周期的奇数倍时, $V(\bar{y}_{sy})$ 很小, 而这些估计量很大. 这说明对于这种模型, 上述估计量都不很适用.

上面讨论的都是针对某一种模型的. 如果对模型不甚了解, 则建议使用 v_2 或 v_0 . 这是因为这两个估计量对于相当广泛的一类实际总体都是普遍适用的.

8.6.2 不等概率系统抽样的情形

在 8.5.2 段中我们曾指出, 不等概率系统抽样对总体总和 Y 的估计仍可用一般不放回系统抽样的 Horvitz-Thompson 估计 \hat{Y}_{HT} . 但它的两种方差估计都不适用于系统样本:

$$\begin{aligned} v_1(\hat{Y}_{HT}) &= \sum_{i=1}^n \frac{1}{\sigma_i^2} y_i^2 + 2 \sum_{i=1}^n \sum_{j>i}^n \frac{(\sigma_{ij} - \sigma_i \sigma_j)}{\sigma_i \sigma_j \sigma_{ij}} y_i y_j, \\ v_2(\hat{Y}_{HT}) &= \sum_{i=1}^n \sum_{j>i}^n \frac{(\sigma_i \sigma_j - \sigma_{ij})}{\sigma_{ij}} \left(\frac{y_i}{\sigma_i} - \frac{y_j}{\sigma_j} \right)^2. \end{aligned}$$

这是因为上述估计量表达式中的分母都含有 σ_{ij} , 而对于系统抽样, σ_{ij} 有可能等于零. 另一个原因是 σ_{ij} 即使不等于 0, 也不易计算 (特别是对 $n > 2$ 的情况). 但 J. N. K. Rao (1962) 证明了当总体单元是随机排列的, 且 $\sigma_i = O(N^{-1})$ 时, 有

$$\sigma_{ij} = \frac{n-1}{n} \sigma_i \sigma_j + \frac{n-1}{n^2} (\sigma_i^2 \sigma_j + \sigma_i \sigma_j^2) - \frac{n-1}{n^3} \sigma_i \sigma_j \sum_{u=1}^N \sigma_u^2 + O(N^{-3}). \quad (8.82)$$

若用上式的近似值替换 σ_{ij} 代入 $v_2(\hat{Y}_{HT})$ (即 \hat{Y}_{HT} 的 Yates-Grundy-Sen 估计), 则可得到一个较为理想的方差估计:

$$v_0 = \frac{1}{n-1} \sum_{i=1}^n \sum_{j>i}^n \left(1 - \sigma_i - \sigma_j + \sum_{u=1}^N \frac{\sigma_u^2}{n} \right) \left(\frac{y_i}{\sigma_i} - \frac{y_j}{\sigma_j} \right)^2 + O(N). \quad (8.83)$$

在上式中, 若 $\sigma_u = \frac{n}{N}$, 即等概率系统抽样的情形, 上式相当于 8.6.1

段中对总体均值 \bar{Y} 的估计量 \bar{y}_{HT} 的方差估计(相差 N^2 倍) v_1 .

如果我们将样本作为放回的 PPS 样本处理, 则可得到另一个方差估计:

$$v_{10} = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{ny_i}{\pi_i} - \hat{P}_{HT} \right)^2. \quad (8.84)$$

由于实际抽样是不放回的, 因此上式“高估”了方差, 还必须考虑有限总体校正系数 $1-f$. 对于此种情形, 一个有用而简单的 f 的估计为:

$$\hat{f} = \sum_{i=1}^n \frac{\pi_i}{n}. \quad (8.85)$$

因此我们可得到另一个方差估计量:

$$v_{11} = (1-\hat{f})v_{10} = \frac{1 - \sum_{i=1}^n \pi_i/n}{n(n-1)} \sum_{i=1}^n \left(\frac{ny_i}{\pi_i} - \hat{P}_{HT} \right)^2. \quad (8.86)$$

与 v_2 及 v_8 的考虑类似, 用相邻样本单元(的加权值)差值的平方和来表示方差, 则有

$$v_{12} = \frac{1-\hat{f}}{n} \sum_{i=2}^n \left(\frac{ny_i}{\pi_i} - \frac{ny_{i-1}}{\pi_{i-1}} \right)^2 / [2(n-1)], \quad (8.87)$$

$$v_{13} = \frac{1-\hat{f}}{n} \sum_{i=1}^{n/2} \left(\frac{ny_{2i}}{\pi_{2i}} - \frac{ny_{2i-1}}{\pi_{2i-1}} \right)^2 / n. \quad (8.88)$$

与此同理, 可构造相当于 v_4 、 v_5 、 v_6 的估计量.

若将样本随机分组成 m 个系统子样本, 每个包含 n/m (设为整数) 个样本单元, 令

$$\hat{P}_\alpha = \frac{m}{n} \sum_{i=1}^{n/m} \frac{ny_{\alpha i}}{\pi_i} \quad (8.89)$$

是第 α 个子样本对 Y 的 HT 估计, 则有如下的估计量

$$v_{14} = \frac{1}{m(m-1)} \sum_{\alpha=1}^m (\hat{P}_\alpha - \hat{P}_{HT})^2. \quad (8.90)$$

如果考虑总体是从某个超总体随机产生的, 则根据该超总体的模型可构造相应的估计量. 不过这类估计量所包含的计算量较大, 具体方法见 Bartley(1962).

上述方差估计量的理论性质目前所知甚少. Wolter(1985) 对这些估计量作了大量的模拟研究与比较, 根据这些模拟研究, 对于随机(或近似随机)排列的总体, v_9 、 v_{10} 与 v_{11} 的效果都比较好. 考虑到 v_9 包含较多的计算, 因此 v_{10} 与 v_{11} 更为人们所乐意采用. 对于具有某种(例如线性)趋势的总体, v_{12} 、 v_{13} 较好. 对于小样本情形, v_{12} 更为适宜. 而与等概率系统抽样中的 v_7 一样, 通常 v_{14} 的性质不太理想, 一般不推荐使用.

第 9 章

复杂样本方差估计的一般方法

§ 9.1 引言

随着社会的发展, 抽样调查的应用日益广泛, 相应的抽样调查的理论研究当然也得到发展。像统计推断一样, 在分析和解释抽样所得的数据资料时, 抽样调查工作者面临两个必须解决的问题: 一为构造一个合适的统计量以对感兴趣的总体指标(参数)作出估计; 二为对所作出的估计进行一定的评价, 即刻划该估计量的精确程度。最通用的关于精确度的测度是调查估计量的方差。一般情况, 估计量的方差是未知的但必须从调查资料本身得到它的估计。

显然, 调查统计量的方差受到统计量本身的形式以及抽样方案的设计特性这两个因素的影响, 因此我们很自然地认为估计量的方差是关于统计量形式及抽样方案的函数。在抽样方案比较简单情况, 例如简单随机抽样、分层随机抽样、二阶抽样、整群抽样等, 而且调查统计量取简单的关于观测值的线性函数形式, 对这种较简单形式的方差估计在本书的前几章中已分别有所介绍。但在实际问题中, 所应用的抽样方案并非简单的一种形式, 通常是几种抽样方法的有机组合, 所采用的估计量也不一定局限于简单估计形式, 可能是诸如比估计、回归估计或其他更复杂的形式。对这类复杂样本以及更一般的估计量, 我们也需要估计其方差。在本章中, 我们将把注意力集中于对这类复杂样本方差估计的一般方法研究上。

9.1.1 复杂样本调查

复杂样本就是从一个复杂抽样调查所得到的样本。关于复杂抽样调查常常从如下几个角度出发考虑:

一、抽样方案的复杂程度

复杂抽样调查常包括一些抽样方案的特性, 诸如: 分层、多阶抽样、不

等概率抽样、双重抽样及多框架等等。

二、调查估计量的复杂程度

复杂的调查估计量常包括那些非线性估计量,例如比估计或回归估计等。有时候,我们需要对某些情况出现而作出一些调整,例如调查中无回答的情况(参见下章)、或者出现过份的“突出值”情况等等,这样的调整当然增加了调查统计量的复杂程度。

三、感兴趣变量或指标的多重性

在大多数有关抽样的教科书中以及本书的前几章中,人们常常一次仅考虑一个指标(参数),而复杂抽样调查常包含数十甚至数百个感兴趣的指标(参数)。

四、调查资料的描述性与分析性用途

复杂抽样调查不仅仅关系到总体的若干指标,除了描述性的目的之外,它还包括分析性的目的。这样可以分析原因,找出总体中的某些关系,进而建立有关的数学模型。

五、调查的规模、范围与深度

如果调查涉及到成千上万个体,需要大规模的组织工作,这样的抽样调查自然是复杂的。

当然,调查的复杂与否并不完全从上述角度清晰地划分。有些调查从某些角度来看可能是复杂的,但从另外的角度来看也许并不复杂。

9.1.2 方法概述

对于一个复杂样本,如何为调查估计量选择一个合适的、近似的方差估计?这实在是一个颇为困难的问题,因为它涉及到方差估计的精度、所花费用(包括时间)的多少、操作的简便性等等。调查统计工作者必须对这些问题给予考虑并且在它们之间作出一定的权衡。

本章将介绍一些非标准的方差估计方法,用这些方法所得到的估计量不一定是无偏估计,但是它们充分变通地迎合了复杂抽样的大多数特性。

Jackknife方法与Bootstrap方法是建立在再抽样理论上的构造方差估计的两种近代统计方法。利用再抽样技巧可以将原来的总体进行复制,在复制的总体中,可以使用原来的抽样办法再复制抽样样本及构造同样结构的有关指标(参数)的统计量。由于复制的总体及统计量是原有总体及统计量的一个缩影,而在复制的模型中,包括统计量的均值、方差

等特性在内的几乎一切为我们所关心的指标均可以通过计算得到——尽管有时候某些计算相当繁复，但是从理论角度来看，由于复制总体为已知，总可以计算出来。——于是，复制模型中统计量的方差作为原来的方差估计的一种替代是顺理成章的。

利用复制技巧对复杂抽样调查实施方差估计的另一种最基本的方法之一是随机组方法，也是最早得到发展的方差估计方法。其实质是按一定的抽样方案从母体中抽取若干组样本，对于每一组样本建立有关指标(参数)的相同形式估计量。这些估计量之间的离散程度提供了基于联合抽样样本所建立的估计量方差的估计。在本章关于 Jackknife 方法这一节中可以看到，关于联合样本所建立的估计量的随机组方差估计实质上是再抽样方法中当再抽样容量大小为随机时“不完全和”的计算形式。

平衡半样本方法是又一种复制技巧的成果，它将(各层中)随机组数减为两个以提高方差估计计算的效率，但是它与随机组方法有所区别。本章将专设一节加以讨论。

区别于上述“复制”估计量技巧的方差估计方法，我们主要介绍 Taylor 级数法。所谓 Taylor 级数法，实质上是一种线性化方法。在抽样调查中人们会遇到一些非线性估计量，比如比估计、回归系数估计等，大多数非线性估计量可以近似地看作为某线性估计，于是利用 Taylor 级数展开的手法可以得到近似的方差估计。在第4章中已讨论过这一技术，在这一章中继续将这一方法予以深化。至于鞍点逼近方法，目前主要用于所考虑的指标为连续变量或者总体元素充分多时，可以近似地将该指标视作连续变量的情况。从抽样所得的经验分布函数出发，利用鞍点方程解得鞍点，然后借助于数学中鞍点逼近的技巧近似地获得统计量分布函数(或分布密度)的估计，这样就可以获得待估参数的置信区间或其他有关信息。这种方法的优点在于当抽样大小 n 比较小时仍然比较精确，从分布拟合的角度优于通常所采用的正态近似。对于我们需要用少量的抽样以对社会经济某些现象作出快速推断来说，无疑提供了一种有价值的方法。

§ 9.2 随机组方法

9.2.1 基本思想与方法

数理统计的常识告诉我们：如果 X_1, X_2, \dots, X_n 为来自同一总体的

互不相关变量, 那么关于变量 X 的方差可以用 $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 作为估计而且具有无偏性. 其中 \bar{X} 是所有 X_i 的平均值. 显然, $\frac{1}{n(n-1)} \times \sum_{i=1}^n (X_i - \bar{X})^2$ 就成了统计量 \bar{X} 的方差的无偏估计.

上述基本常识启示我们: 如果从有限总体的一个样本得到有关指标 (或参数 θ 的一个估计量, 不妨假设为 $\hat{\theta}_1$, 那末重复同样的抽样方法以及构造同样形式的估计量若干次, 可以得到若干个 (比如 k 个) 关于 θ 的估计量 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$, 且基于 k 组联合样本可以产生一个新的关于 θ 的联合估计, 记作

$$\hat{\theta} = \frac{1}{k} \sum_{i=1}^k \hat{\theta}_i.$$

显然 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ 构成了来自 $\hat{\theta}_1$ 所属总体的一系列随机观测值, 于是 $\hat{\theta}$ 的方差估计可以取作

$$v(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{i=1}^k (\hat{\theta}_i - \hat{\theta})^2,$$

它是参数 $V(\hat{\theta})$ 的无偏估计. 构成这些估计量 $\hat{\theta}_i (i=1, 2, \dots, k)$ 的 k 组抽样, 即为来自总体的 k 个随机组.

方差的所谓随机组估计实质上选择来自总体的两个或两个以上组抽样, 一般地, 每组抽样采用相同抽样方案, 对每组抽样分别构造关于参数 θ 的估计量, 利用这些估计量之间差的平方计算基于所有样本联合估计量的方差.

随机组方法有两种基本形式: 一为随机组之间互为独立; 二为随机组之间存在某种类型的相依性.

9.2.2 独立随机组

以一定的抽样方式取第一组样本 s_1 , 然后放回总体, 再按原来的抽样方式取第二组样本 s_2 , 再放回总体; \dots ; 重复上述步骤 k 次, 可以得到 k 组随机样本: s_1, s_2, \dots, s_k . 对于每一随机组样本, 以某种形式确定 θ 的估计, 于是得到 k 个独立的关于 θ 的估计, 记作 $\hat{\theta}_\alpha (\alpha=1, 2, \dots, k)$.

描述方差的随机组估计的主要结论陈述如下:

定理 9.1 设 $\hat{\theta}_1, \dots, \hat{\theta}_k$ 为具有共同期望 $E(\hat{\theta}_1) = \mu$ 的 k 个互不相关的随机变量, 定义 $\hat{\theta}$ 为

$$\hat{\theta} = \sum_{\alpha=1}^k \hat{\theta}_\alpha / k.$$

那末 $E(\hat{\theta}) = \mu$ 且 $V(\hat{\theta})$ 的无偏估计为

$$v(\hat{\theta}) = \sum_{\alpha=1}^k (\hat{\theta}_{\alpha} - \hat{\theta})^2 / k(k-1). \quad (9.1)$$

证明 $E(\hat{\theta}) = \mu$ 是显然的事实. 若记

$$v(\hat{\theta}) = \left[\sum_{\alpha=1}^k \hat{\theta}_{\alpha}^2 - k\hat{\theta}^2 \right] / k(k-1),$$

则

$$\begin{aligned} E\{v(\hat{\theta})\} &= \left[\sum_{\alpha=1}^k (V(\hat{\theta}_{\alpha}) + \mu^2) - k(V(\hat{\theta}^2) + \mu^2) \right] / k(k-1) \\ &= (k^2 - k)V(\hat{\theta}) / k(k-1) = V(\hat{\theta}). \quad \blacksquare \end{aligned}$$

注意到定理 9.1 中并没有要求随机变量 $\hat{\theta}_{\alpha}$ 的方差相等, 这意味着随机组样本可以用不同的抽样方式取得, 而且 $\hat{\theta}_{\alpha}$ 也可以取不同的函数形式, 只需要这些 $\hat{\theta}_{\alpha}$ 互不相关且具有共同的期望, 定理的结论依然成立. 因此利用 (9.1) 对 $\hat{\theta}$ 的方差进行估计是顺乎自然的事情. 关于 θ 的进一步推断常常离不开关于 θ 的置信区间这一概念. 众所周知, 在数理统计领域有如下著名的结论回答了这个问题:

定理 9.2 假设 $\hat{\theta}_1, \dots, \hat{\theta}_k \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$, 那末

(1) 统计量

$$z = (\hat{\theta} - \theta) / \sqrt{\sigma^2/k} \sim N(0, 1).$$

(2) 统计量

$$t = (\hat{\theta} - \theta) / \sqrt{v(\hat{\theta})} \sim t(k-1).$$

利用这个结论, 假如 $\hat{\theta}$ 的方差基本上已知且无误差, 或者 k 相当大的话, 那末 θ 的 $(1-\alpha)$ 置信区间为

$$(\hat{\theta} - u_{\alpha} \sqrt{v(\hat{\theta})}, \hat{\theta} + u_{\alpha} \sqrt{v(\hat{\theta})}),$$

其中 u_{α} 为标准正态分布的双侧 α 分位点; 而当 $\hat{\theta}$ 的方差未知, 或者 k 并不十分大时, 置信区间取作

$$(\hat{\theta} - t_{k-1, \alpha} \sqrt{v(\hat{\theta})}, \hat{\theta} + t_{k-1, \alpha} \sqrt{v(\hat{\theta})}),$$

其中 $t_{k-1, \alpha}$ 是分布 $t(k-1)$ 的双侧 α 分位点.

很明显, 定理 9.2 的条件强于定理 9.1 的条件, 而且在有限总体抽样中很难严格地满足. 但是独立随机组抽样以及在每一个随机组中用同样的方法构造 θ 的估计量, 就常常在方法上保证了 $\hat{\theta}_1, \dots, \hat{\theta}_k$ 的独立同分布性. 至于 $\hat{\theta}_{\alpha}$ 的正态性假设通常在有限总体抽样中根本无法成立, 然而抽样调查的渐近理论常常在大样本情况下保证了 $\hat{\theta}_{\alpha}$ 的近似正态性. 至于 $\hat{\theta}_{\alpha}$ 是否期望为 θ 的问题, 对于非线性估计量来说, 常存在非零偏倚, 好

在大样本量的抽样保证了这样的偏倚常常显得微不足道。

概括起来, 定理 9.1 与定理 9.2 使得在独立随机组抽样方法的很多场合, 可以得到 $\hat{\theta}$ 的方差的无偏估计, 并且视 k 的大小, 利用正态理论或 t 分布理论获得 θ 的置信区间。

由此, 不难想象随机组抽样方法的许多重要应用可能在于非线性统计量。如果基于所有 k 个随机组的联合样本, 用构造 $\hat{\theta}_\alpha$ 的方式同样构造 θ 的估计量 $\hat{\theta}$ 而不是简单地将 $\hat{\theta}_\alpha$ 平均而得 $\bar{\hat{\theta}}$, 似乎是自然(且或许更有效)的考虑。倘若 $\hat{\theta}_\alpha$ 是线性估计的话, $\bar{\hat{\theta}}$ 与 $\hat{\theta}$ 是相同的。但是对于非线性估计量而言, 它们一般并不相等。以下例子很清楚地说明这个问题。

例 9.1 假如希望估计两个调查指标总和 Y 与 X 之比 $\theta = Y/X$, 设 $\hat{Y}_\alpha, \hat{X}_\alpha (\alpha = 1, 2, \dots, k)$ 分别表示第 α 随机组中关于 Y 与 X 的估计, 实用中它们常为线性无偏估计。于是

$$\begin{aligned}\hat{\theta}_\alpha &= \hat{Y}_\alpha / \hat{X}_\alpha, \\ \bar{\hat{\theta}} &= \frac{1}{k} \sum_{\alpha=1}^k (\hat{Y}_\alpha / \hat{X}_\alpha), \\ \hat{\theta} &= \sum_{\alpha=1}^k \hat{Y}_\alpha / \sum_{\alpha=1}^k \hat{X}_\alpha.\end{aligned}$$

显然, 一般地 $\bar{\hat{\theta}} \neq \hat{\theta}$ 。

对于 $\hat{\theta}$ 的方差在实际中有两种随机组估计:

$$v_1(\hat{\theta}) = \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \bar{\hat{\theta}})^2 / k(k-1), \quad (9.2)$$

$$v_2(\hat{\theta}) = \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 / k(k-1), \quad (9.3)$$

若 $\hat{\theta}_\alpha$ 为线性形式, 由于 $\bar{\hat{\theta}} = \hat{\theta}$ 从而 $v_1 = v_2$ 。但是对于非线性估计量来说, 有

$$\sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 = \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \bar{\hat{\theta}})^2 + k(\bar{\hat{\theta}} - \hat{\theta})^2.$$

当然有

$$v_1(\hat{\theta}) \leq v_2(\hat{\theta})$$

成立。注意: 我们希望估计的是 θ 的方差而不是得到一个“最小”估计, 故从稳妥的角度出发, 一般地宁取 $v_2(\hat{\theta})$, 因为它体现了各分组中的 $\hat{\theta}_\alpha$ 关于 $\hat{\theta}$ 的差平方的平均。在复杂抽样中, 由于样本量一般较大, $E(\bar{\hat{\theta}} - \hat{\theta})^2$ 通常显得微不足道, 因此在 v_1 与 v_2 之间没有太大的差异。倘若在 v_1 与 v_2 (或者 $\bar{\hat{\theta}}$ 与 $\hat{\theta}$) 之间存在显著差异的话, 则要么说明在计算中发生了误差, 要么表明这是由于抽样容量偏小而引起的偏倚。

对 v_1 与 v_2 , 到底推荐以哪个为 $\hat{\theta}$ 的方差估计是个难确定的问题, 到

底哪一个是 $V(\hat{\theta})$ 的较精确估计, 仍然是个需要探讨的课题.

9.2.3 非独立随机组

在实际应用中, 很少会进行一系列的独立随机组抽样, 最经常的办法是采用某种不放回形式整体地选择调查样本. 此时的随机组只能采用将这些样本随机地划分为 k 组, 然后在每组中计算估计量, 并采用 (9.1) 形式的方差估计公式. 显然这种划分随机组的方法使得各个 $\hat{\theta}_\alpha$ 之间不再互不相关, 定理 9.1 的结论不再严格地成立.

如何将原始样本随机地划分为 k 组呢? 一个最基本的原则是使每个随机组具有与原始样本一样的抽样结构, 或者说, 由于划分的随机性, 使每个随机组均可看作原始样本的一个缩影. 这在简单不放回或不放回 PPS 形式抽取 n 个单阶样本的情况不难办到. 我们只需从原始样本中不放回地随机抽取 $m = [n/k]$ 个样本单元作为第一个随机组, 从余下的 $n - m$ 个原始样本单元中再不放回地随机抽取 m 个作为第二个随机组, 继续这种做法直到抽完. 假如 n/k 不是整数, 即 $n = km + q$ ($0 < q < k$), 那末有两种处理方式: 要么将最后这 q 个样本排除出 k 个随机组之外, 这样估计量将损失一定的信息; 要么将它们逐一加入前面 q 个随机组. 对于较复杂一些的抽样则要小心一些. 例如, 在多阶抽样模型, 随机组的形式是将最基本的群 (即对相同初级抽样单元所选取的样本的集合) 划分为 k 组而得到. 这样的划分原则依赖于首阶抽样方案的特性; 对于分层抽样, 有两种选择: 倘若希望估计某层内的方差, 那么按照在该层内抽样方案的特性在该层划分随机组; 倘若希望估计包括所有层在内的总方差, 那么每一个随机组必须是又一个分层抽样, 即对每一层中抽得的样本按照在该层中原来的抽样方案随机地划分 k 组, 然后在各层抽样中任意各取一个随机组形成总抽样的一个随机组, \dots , 等等.

在非独立随机组情况下, 关于总体参数 θ 的估计方法一般与独立随机组的情况一样. 设 $\hat{\theta}$ 表示从原样本中计算而得的 θ 的估计量, $\hat{\theta}_\alpha$ 表示第 α 随机组中 θ 的估计量, 且 $\hat{\theta} = \sum_{\alpha=1}^k \hat{\theta}_\alpha / k$. $V(\hat{\theta})$ 的随机组估计为

$$v(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2. \quad (9.4)$$

而对于 $\hat{\theta}$ 的方差通常有两种估计:

$$v_1(\hat{\theta}) = v(\hat{\theta}), \quad (9.5)$$

$$v_2(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2. \quad (9.6)$$

出于同样理由, 为稳妥起见, 我们有时宁取 v_2 而不取 v_1 .

由于随机组估计 $\hat{\theta}_\alpha$ 相互之间不再是独立的, 通常 $v(\hat{\theta})$ 不再是 $V(\hat{\theta})$ 的无偏估计, $v(\hat{\theta})$ 的一些性质可由下述定理描述:

定理 9.3 设 $E(\hat{\theta}_\alpha) = \mu_\alpha$, μ_α 不必等于 θ , 则

$$E(\hat{\theta}) = \sum_{\alpha=1}^k \mu_\alpha / k \triangleq \bar{\mu},$$

$$\begin{aligned} \text{且} \quad E\{v(\hat{\theta})\} &= V(\hat{\theta}) + \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\mu_\alpha - \bar{\mu})^2 \\ &\quad - 2 \sum_{\alpha=1}^k \sum_{\beta>\alpha}^k \text{Cov}(\hat{\theta}_\alpha, \hat{\theta}_\beta) / k(k-1). \end{aligned}$$

如果每个随机组具有相同大小, 那末

$$\mu_\alpha = \bar{\mu} (\alpha = 1, 2, \dots, k),$$

$$E(\hat{\theta}) = \bar{\mu},$$

$$\text{且} \quad E\{v(\hat{\theta})\} = V(\hat{\theta}) - \text{Cov}(\hat{\theta}_1, \hat{\theta}_2).$$

证明 $E(\hat{\theta}) = \bar{\mu}$ 是显然的事实. 将方差的随机组估计改写为

$$v(\hat{\theta}) = \hat{\theta}^2 - 2 \sum_{\alpha=1}^k \sum_{\beta>\alpha}^k \hat{\theta}_\alpha \hat{\theta}_\beta / k(k-1),$$

$$\text{由} \quad E(\hat{\theta}^2) = V(\hat{\theta}) + \bar{\mu}^2$$

$$\text{及} \quad E(\hat{\theta}_\alpha \hat{\theta}_\beta) = \text{Cov}(\hat{\theta}_\alpha, \hat{\theta}_\beta) + \mu_\alpha \mu_\beta,$$

不难得到定理的结论. ■

9.2.4 随机组数 k 的选择

正如统计推断所处理的那样, 对于 θ 的估计量 $\hat{\theta}$ 的方差, 首先需要得到它的估计. 这不仅在分析资料时显得非常重要, 而且在设计调查方案时估计量 $\hat{\theta}$ 的方差也显得相当重要. 因为调查统计工作者可以利用 $\hat{\theta}$ 的方差估计而设法将调查方案最佳化并选择足够大的样本以产生关于 $\hat{\theta}$ 的精确程度的理想水平. 第二个重要问题在于 $\hat{\theta}$ 的方差估计的精度. 我们已经指出利用独立或非独立随机组方法至少部分回答了上述第一个重要性, 紧接着关心的一个问题自然是该方差估计的精度. 为达到较理想的精度, 人们自然要问: “需要划分多少个随机组?” 即 k 究竟选择什么样的整数为最佳?

谈到随机组方差估计 $v(\hat{\theta})$ 的质量评估, 毫无疑问会想到 $v(\hat{\theta})$ 的方差

$V\{v(\hat{\theta})\}$. 然而此时我们一般并不太关心 $v(\hat{\theta})$ 的置信区间, 因此纯粹地再估计 $V\{v(\hat{\theta})\}$ 意义不大. 况且一个变量的稳定与否不单纯考虑它的方差大小, 还要顾及到关于变量的平均值的相对大小加以考虑, 于是产生了一般的 OV 准则, 即考虑 $v(\hat{\theta})$ 的变异系数:

$$OV\{v(\hat{\theta})\} = [V\{v(\hat{\theta})\}]^{1/2}/V(\hat{\theta}).$$

另一个准则是考虑 θ 的置信区间:

$$(\hat{\theta} - c\{v(\hat{\theta})\}^{1/2}, \hat{\theta} + c\{v(\hat{\theta})\}^{1/2})$$

的大小(其中 c 为某正常数), 或者考虑它包含 θ 的覆盖概率. 当然还有一些其他的评估准则, 这要视它在统计分析中的用途而定, 本书不准备一一加以论述了.

对于主要的 OV 准则, 由下述定理开始探讨:

定理 9.4 设 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ 为独立同分布变量, 而 $v(\hat{\theta})$ 如公式(9.1)所定义, 那末

$$OV\{v(\hat{\theta})\} = \left\{ \frac{\beta_4(\hat{\theta}_1) - (k-3)/(k-1)}{k} \right\}^{1/2}, \quad (9.7)$$

其中
$$\beta_4(\hat{\theta}_1) = \frac{E\{(\hat{\theta}_1 - \mu)^4\}}{[E\{(\hat{\theta}_1 - \mu)^2\}]^2}, \quad \mu = E(\hat{\theta}_1).$$

证明 由 $\hat{\theta}_\alpha (\alpha = 1, 2, \dots, k)$ 的独立性, 有

$$E\{v^2(\hat{\theta})\} = \frac{1}{k^4} \sum_{\alpha=1}^k \kappa_4(\hat{\theta}_\alpha) + \frac{2}{k^4} \left(1 + \frac{2}{(k-1)^2}\right) \sum_{\alpha=1}^k \sum_{\beta \neq \alpha}^k \kappa_2(\hat{\theta}_\alpha) \kappa_2(\hat{\theta}_\beta),$$

其中
$$\begin{aligned} \kappa_4(\hat{\theta}_\alpha) &= E\{(\hat{\theta}_\alpha - \mu)^4\}; \\ \kappa_2(\hat{\theta}_\alpha) &= E\{(\hat{\theta}_\alpha - \mu)^2\}. \end{aligned}$$

注意到 $\hat{\theta}_\alpha$ 的同分布特性, 故

$$V\{v(\hat{\theta})\} = \frac{1}{k^3} \kappa_4(\hat{\theta}_1) + \frac{k-1}{k^3} \cdot \frac{k^2-2k+3}{(k-1)^2} \kappa_2^2(\hat{\theta}_1) - E^2\{v(\hat{\theta})\}.$$

由变异系数定义即得定理结论. \blacksquare

定理 9.4 实质上告诉我们, 独立随机组方差估计的 OV 依赖于峰态 $\beta_4(\hat{\theta}_1)$ 及组数 k 这两个因素. 如果 k 小, OV 则大, 从而方差估计具较差精度. 如果 $\hat{\theta}_\alpha$ 的频率曲线在中心附近及在尾部具有“超越”量, 峰态 $\beta_4(\hat{\theta}_1)$ 就大, 而方差估计的精度就差. 假如 k 比较大, 则 OV^2 近似地反比于随机组数 k :

$$OV^2\{v(\hat{\theta})\} \approx \frac{\beta_4(\hat{\theta}_1) - 1}{k}.$$

于是, 方差的随机组估计的精度非但依赖于组数 k , 而且与 $\hat{\theta}_\alpha$ 的分布(从

而与 $\beta_4(\hat{\theta}_\alpha)$ 有密切关系, 也就是说, 与 $\hat{\theta}_\alpha$ 的构造形式和样本抽取的方式均有关系. 假如 $\hat{\theta}$ 取为样本的某平均形式, 我们可以容易地计算出 $\beta_4(\hat{\theta}_\alpha)$; 设 n 恰为随机组数 k 的整数倍, 即每个随机组含 $m = n/k$ 个样本单元, 若原抽样方式为放回的简单随机抽样, $\hat{\theta}_\alpha$ 取作第 α 组的样本均值, 显然

$$\hat{\theta} = \frac{1}{k} \sum_{\alpha=1}^k \hat{\theta}_\alpha = \frac{1}{n} \sum_{i=1}^n y_i.$$

此时

$$\beta_4(\hat{\theta}_1) = \beta_4/m + 3(m-1)/m,$$

$$\beta_4 = \left[\sum_{i=1}^k (Y_i - Y)^2 / N \right] / \left\{ \sum_{i=1}^k (Y_i - Y)^2 / N \right\}^2.$$

而若原始样本为在放回情况下的 PPS 样本, 今取 $\hat{\theta}_\alpha = \frac{1}{m} \sum_{i=1}^m y_i/z_i$ 表示基于第 α 随机组的总体总和的估计量, 那么

$$\hat{\theta} = \frac{1}{k} \sum_{\alpha=1}^k \hat{\theta}_\alpha = \frac{1}{n} \sum_{i=1}^n y_i/z_i.$$

此时

$$\beta_4(\hat{\theta}_1) = \beta_4/m + 3(m-1)/m,$$

$$\beta_4 = \frac{\sum_{i=1}^k (T_i - T)^2 / N}{\left\{ \sum_{i=1}^k (T_i - T)^2 / N \right\}^2},$$

其中

$$T_i = Y_i/z_i.$$

这两种特殊的形式蕴含了这样一个事实. 常见的实际情况是 $\beta_4(\hat{\theta}_1)$ 基本上具有 $\frac{a}{m} + b$ 的形式, 其中 a 、 b 为常数. 当 m 从 1 开始增加时, $\beta_4(\hat{\theta}_1)$ 明显地减少, 然而随着 m 越来越大, 峰度 $\beta_4(\hat{\theta}_1)$ 的递减显得越来越不重要, 它抵消不了相应减少的 k 所带来的影响, 因此组数 k 比起组内样本量 m 来, 对 $OV\{v(\hat{\theta})\}$ 的减小与方差估计精度的提高, 具有较大的影响.

定理 9.4 的结果可以在不放回抽样这样更普遍的场所, 尤其是在总体很大而抽样比较小的场合近似地认可, 这在 Hansen、Hurwitz 与 Madow (1953) 的书中有阐述.

现在问题回到随机组数 k 的选择, 在前面讨论了从精度出发, 我们乐意取 k 尽可能地大, 然而, 增加 k 就意味着增加计算工作量与成本, 于是 k 的最佳值应该是成本与精度的平衡, 这个问题自然地随着调查的不同而变化.

假如在某种场合, 调查的目的仅仅在于得到关于总体某指标的一个

粗糙概念, 那么对成本方面的考虑可能要超过对精度方面的考虑, 从而对 k 的最佳值不妨取小一些. 另一方面, 假如重要的决策主要基于调查的结果, 此时对精度的要求就要超过对成本方面的考虑, 从而一般应取较大的 k 值.

§ 9.3 Jackknife 方法与 Bootstrap 方法

9.3.1 Jackknife 的基本思想与方法

在第4章中提到的 Quenouille 的 Jackknife 方法原本是在时间序列分析中用于估计量的纠偏. 假如 θ 的估计量为 $\hat{\theta}_n(x_1, x_2, \dots, x_n) \triangleq \hat{\theta}_n$, 在样本中舍弃第 j 个观察值后用同样方式得到 θ 的估计量为 $\hat{\theta}_{-j}$, 构造所谓虚拟值(Pseudovalues):

$$\hat{\theta}_{n,j} = n\hat{\theta}_n - (n-1)\hat{\theta}_{-j} \quad (j=1, 2, \dots, n). \quad (9.8)$$

所有 n 个虚拟值的平均值称为 $\hat{\theta}_n$ 的 Jackknife 形式:

$$\hat{\theta}_J = \frac{1}{n} \sum_{j=1}^n \hat{\theta}_{n,j}. \quad (9.9)$$

如果 $E(\hat{\theta}_n) = \theta + a/n$ (a 与 n 无关), 那末容易验证 $\hat{\theta}_J$ 为 θ 的无偏估计. 倘若 $\hat{\theta}_n$ 具有关于 $1/n$ 的更高阶偏倚, 我们可以用高阶 Jackknife 方法处理.

众所周知, 抽样推断所选择的统计量 $\hat{\theta}_n$ 通常包含各个随机观测值提供的关于 θ 的全部信息. 相应地, $\hat{\theta}_{-j}$ 就包含了除去第 j 个观测值之外其他随机观测值提供的关于 θ 的信息. (9.8) 式实质上蕴含了这样一个事实: 虚拟值既然是从 $\hat{\theta}_n$ 中关于 θ 的信息中剔除了 $\hat{\theta}_{-j}$ 中关于 θ 的信息, 从而虚拟值可以看作几乎仅仅包含 X_j 所提供的关于 θ 的信息. 例如对于样本均值来说, 容易验证, 它的第 j 个虚拟值恰为 X_j . 假如这些 X_j 是互相独立的随机观测值, 那末这些虚拟值之间自然就存在着某种程度的独立性. 基于这种思想, Tukey(1958) 在一篇很短但却很著名的摘要中提出了如下猜想:

假如 $\hat{\theta}_n$ 为基于独立同分布变量 X_1, X_2, \dots, X_n 的关于 θ 的估计量, 相应的虚拟值及 Jackknife 估计如(9.8)与(9.9)式所定义的, 那末

(1) 虚拟值 $\hat{\theta}_{n,j}$ ($j=1, 2, \dots, n$) 可以近似地看作为独立同分布的随机变量.

(2) 基于假设(1), 统计量

$$\frac{\sqrt{n}(\hat{\theta}_J - \hat{\theta})}{V_{J1}} \xrightarrow{\mathcal{D}} t(n-1), \quad (9.10)$$

式中 $\xrightarrow{\mathcal{D}}$ 表示以分布收敛, 且

$$V_{J1} = \frac{1}{n-1} \sum_{j=1}^n (\hat{\theta}_{n,j} - \hat{\theta}_J)^2. \quad (9.11)$$

显然, Tukey 猜想将 Jackknife 方法的纠偏作用扩展到用以构造 θ 的置信区间以及获取估计量的方差估计. 从此 Jackknife 方法显示出其强盛的生命力. 近几十年来, 不少统计工作者围绕着何种情况何种统计量满足 Tukey 猜想这个课题进行了大量研究, 出现了不少文献(参见施锡铨(1987); Miller, R.G., Jr. The Jackknife: A Review, Biometrika, (1974), 61, 1~15). 在对有限总体抽样中应用 Jackknife 方法也许首推 Durbin(1959), 他在比估计的 Jackknife 方差估计问题上获得了成功.

9.3.2 有限总体的 Jackknife 方差估计

在有限总体应用中, 我们常采用 Jackknife 的更一般的形式, 它与随机组方法有着一定关系:

将样本分成 k 个随机组(假定 $n = km$, m 为整数), 这些组当然可以分为独立与不独立的两种情况. 以 $\hat{\theta}$ 表示基于原始样本的关于 θ 的估计量, 而 $\hat{\theta}_{-\alpha}$ 则表示舍弃第 α 组观测值后关于 θ 的具有 $\hat{\theta}$ 同样结构的估计量, 不难得到虚拟值为:

$$\hat{\theta}_{\alpha} = k\hat{\theta} - (k-1)\hat{\theta}_{-\alpha} \quad (\alpha = 1, 2, \dots, k),$$

于是 $\hat{\theta}$ 的 Jackknife 形式为

$$J(\hat{\theta}) = \hat{\bar{\theta}} = \sum_{\alpha=1}^k \hat{\theta}_{\alpha} / k. \quad (9.12)$$

其相应的方差估计为

$$v(\hat{\bar{\theta}}) = \sum_{\alpha=1}^k (\hat{\theta}_{\alpha} - \hat{\bar{\theta}})^2 / k(k-1). \quad (9.13)$$

在本节开头引入的 Jackknife 方法无非是 $k=n$ 的特殊情况.

以下讨论有限总体 Jackknife 方差估计的若干情况:

一、放回简单随机抽样

设总体的单元为 Y_1, Y_2, \dots, Y_N , 待估参数为总体均值 $\bar{Y} = \sum_{j=1}^N Y_j / N$, 假如从该总体中简单随机有放回的抽取样本 y_1, y_2, \dots, y_n 那末 $\bar{y} = \sum_{i=1}^n y_i / n$ 是 \bar{Y} 的无偏估计, 其方差为 $V(\bar{y}) = \sum_{j=1}^N (Y_j - \bar{Y})^2 / nN$, 该

方差通常具有无偏估计

$$v(\bar{y}) = \sum_{i=1}^n (y_i - \bar{y})^2 / n(n-1). \quad (9.14)$$

应用 Quenouille 的 Jackknife 方法, 若 $n = km$, 令 $\hat{\theta} = \bar{y}$. 于是

$$\hat{\theta} = \sum_{\alpha=1}^k \hat{\theta}_{\alpha} / k = k\bar{y} - (k-1) \sum_{\alpha=1}^k \bar{y}_{-\alpha}, \quad (9.15)$$

其中 $\bar{y}_{-\alpha}$ 表示舍弃第 α 组观测值后得到的样本均值. 容易验证

$$\hat{\theta}_J = \hat{\theta} = \bar{y}, \quad (9.16)$$

$$v(\hat{\theta}) = \frac{k-1}{k} \sum_{\alpha=1}^k (\bar{y}_{-\alpha} - \bar{y})^2, \quad (9.17)$$

当且仅当 $k=n$, $m=1$ 时, (9.14) 与 (9.17) 相等. 但是不难发现如下事实:

$$E\{v(\hat{\theta})\} = V(\hat{\theta}) = V(\bar{y}). \quad (9.18)$$

二、放回 PPS 抽样

假如从总体中进行大小为 n 的放回 PPS 抽样, 各个单元每次被抽取的概率为 z_i ($i = 1, 2, \dots, N$), 对总体总和的通常估计及其估计量方差分别为:

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n y_i / z_i,$$

$$V(\hat{Y}) = \frac{1}{n} \sum_{i=1}^N z_i (Y_i / z_i - Y)^2.$$

(Y 表示总体总和), 对于 $V(\hat{Y})$, 有无偏估计:

$$v(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^n (y_i / z_i - \hat{Y})^2. \quad (9.19)$$

利用 Jackknife 方法, 令 $\hat{\theta} = \hat{Y}$ 且假定 $n = km$, 那么

$$\hat{\theta}_J = \hat{\theta} = k\hat{Y} - (k-1) \sum_{\alpha=1}^k \hat{Y}_{-\alpha}, \quad (9.20)$$

其中 $\hat{Y}_{-\alpha}$ 是舍弃第 α 组后所得的 Y 的估计量. 方差的 Jackknife 估计为

$$v_1(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_{\alpha} - \hat{\theta})^2, \quad (9.21)$$

其中 $\hat{\theta}_{\alpha}$ 是第 α 个虚拟值:

$$\hat{\theta}_{\alpha} = k\hat{Y} - (k-1)\hat{Y}_{-\alpha}.$$

注意到 \hat{Y} 与 \bar{y} 均为 n 个随机变量的平均值形式. 凡统计量具有形式如同观测值或观测值的函数的平均值, 那末只要 $n = km$, 此时不管 k 的取值多少, 容易验证该统计量的 Jackknife 形式必定等于统计量本身, 这是

Jackknife 的重要性质之一. 因而, 在本例中有 $\hat{Y}_j = \hat{\theta}$, 而当 $k=n$ 时, 由于第 i 个虚拟值即为平均值中第 i 个变量, 因此显然有 $v_1(\hat{\theta}) = v(\hat{Y})$ 成立.

三、不放回简单随机抽样

对于通常的简单随机抽样, 即不放回随机抽样样本均值 $\bar{y} = \hat{\theta}$ 用以估计总体均值来说, 它的 Jackknife 形式 $\hat{\theta}_j$, $\hat{\theta} = y$ 是当然的事实, 关键在于此时 \bar{y} 的方差具有形式:

$$V(\bar{y}) = (1-f)S^2/n,$$

其中 $f = n/N$, $S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$, \bar{y} 的方差的无偏估计通常取为

$$v(\bar{y}) = \frac{(1-f)}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2.$$

相应的 Jackknife 方差估计为

$$v_1(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2,$$

$$\hat{\theta}_\alpha = k\bar{y} - (k-1)\bar{y}_{-\alpha} \triangleq \bar{y}_{(\alpha)} \text{ (第 } \alpha \text{ 组观测值的平均值).}$$

利用第二章的计算, 可以得到

$$E\{v_1(\hat{\theta})\} = S^2/n. \quad (9.22)$$

它与 $V(\bar{y})$ 之间存在偏差 fS^2/n , 因而 $v_1(\hat{\theta})$ 不是 $V(\bar{y})$ 的无偏估计, 这是 Jackknife 方差估计在有、无放回两种情况之间的差别. 事实上很明显, 这里的差别主要在于待估方差 $V(\bar{y})$ 之间的差别.

在不放回简单随机抽样时, 如果有限总体校正系数不能忽略, 那么可以采用如下的方差无偏估计:

$$(1-f)v_1(\hat{\theta}).$$

在实用中, 为了达到某种纠偏的目的, 经常对 Jackknife 估计采取一些小小的修整工作, 称之为“修正” Jackknife, 在本节中, 将修正 Jackknife 的虚拟值定义为:

$$\hat{\theta}_\alpha^* = k\hat{\theta} - (k-1)\hat{\theta}_{-\alpha},$$

其中

$$\hat{\theta}_{-\alpha}^* = y + (1-f)^{1/2}(\bar{y}_{-\alpha} - \bar{y}).$$

这样, 就非但有

$$\hat{\theta}^* = \sum_{\alpha=1}^k \hat{\theta}_\alpha^* / k = \bar{y},$$

$$\text{而且 } E\{v_1(\hat{\theta}^*)\} = E\left\{\frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{\theta}_\alpha^* - \hat{\theta}^*)^2\right\} = V(\bar{y}) = (1-f)S^2/n.$$

四、比估计

Jackknife 应用于非线性估计量的方差估计的各类例子中, 最典型的是比估计. 假如需要估计两个总体总和之比 $R = Y/X$, 设 \hat{Y} 、 \hat{X} 分别为 Y 、 X 的估计, 那么, 很自然地采用 $\hat{R} = \hat{Y}/\hat{X}$ 以估计 R . 相应地有

$$\hat{R}_\alpha = \hat{Y}_\alpha / \hat{X}_\alpha.$$

于是得到 Jackknife 虚拟值为:

$$\hat{R}_\alpha = k\hat{R} - (k-1)\hat{R}_\alpha. \quad (9.23)$$

由此得到 Quenouille 的 Jackknife 估计形式为:

$$\hat{\hat{R}} = k^{-1} \sum_{\alpha=1}^k \hat{R}_\alpha. \quad (9.24)$$

对于 $\hat{\hat{R}}$ 或 \hat{R} 的 Jackknife 方差估计则为:

$$v_1(\hat{\hat{R}}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{R}_\alpha - \hat{\hat{R}})^2, \quad (9.25)$$

$$v_2(\hat{R}) = \frac{1}{k(k-1)} \sum_{\alpha=1}^k (\hat{R}_\alpha - \hat{R})^2. \quad (9.26)$$

五、一般情况

假如有 L 个子总体(或 L 层), 第 h 个子总体(或层)中含 N_h 个单元, 从该子总体(层)中有(或无)放回地随机抽取 n_h 个单元($h=1, 2, \dots, L$), 设 \bar{Y}_h 表示该子总体(层)的均值, 而 \bar{y}_h 表示关于 \bar{Y}_h 的估计量. 设我们感兴趣的参数具有如下形式:

$$\theta = g(\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_L).$$

假定 $g(\cdot)$ 是个性质良好、比较光滑的函数, 那末 θ 的一个自然估计当然取作

$$\hat{\theta} = g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L). \quad (9.27)$$

如果 $g(\cdot)$ 光滑到使它至少在 $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_L)$ 的邻域内具有足够阶连续导数的地步, 而 \bar{y}_h 为样本均值因而是 \bar{Y}_h 的无偏估计, 运用 Taylor 展开的方法:

$$\begin{aligned} g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L) &= g(\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_L) \\ &= \sum_{h=1}^L \frac{\partial g}{\partial Y_h} (\bar{y}_h - \bar{Y}_h) + \frac{1}{2} \sum_i \sum_j \frac{\partial^2 g}{\partial \bar{Y}_i \partial \bar{Y}_j} (\bar{y}_i - \bar{Y}_i) (\bar{y}_j - \bar{Y}_j) + \dots, \end{aligned} \quad (9.28)$$

这里及本段其他地方出现的偏导数均在点 $(\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_L)$ 处取值. 注意到各总体(层)的抽样是互为独立的, 近似地有

$$E\{g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)\} \doteq g(\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_L)$$

$$\begin{aligned}
& + \sum_{h=1}^L \frac{1}{2} \cdot \frac{\partial^2 g}{\partial \bar{Y}_h^2} E(\bar{y}_h - \bar{Y}_h)^2 \\
& = \begin{cases} g(\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_L) + \sum_{h=1}^L \frac{1}{n_h} \sigma_h^2 \cdot \frac{1}{2} \frac{\partial^2 g}{\partial \bar{Y}_h^2} \\ \quad \text{(放回抽样情况);} \\ g(\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_L) + \sum_{h=1}^L \frac{N_h - n_h}{(N_h - 1)n_h} \sigma_h^2 \cdot \frac{1}{2} \frac{\partial^2 g}{\partial \bar{Y}_h^2} \\ \quad \text{(不放回抽样情况),} \end{cases} \quad (9.29)
\end{aligned}$$

其中 σ_h^2 为第 h 个子总体(或层)内的方差. 对于 $g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)$ 的方差, 经过计算可以近似地得到

$$\begin{aligned}
V\{g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)\} & \approx \sum_{h=1}^L \left(\frac{\partial g}{\partial \bar{Y}_h} \right)^2 E(\bar{y}_h - \bar{Y}_h)^2 \\
& + \sum_{h=1}^L \left(\frac{\partial g}{\partial \bar{Y}_h} \right) \left(\frac{\partial^2 g}{\partial \bar{Y}_h^2} \right) E(\bar{y}_h - \bar{Y}_h)^3 \\
& = \begin{cases} \sum_{h=1}^L \frac{1}{n_h} \sigma_h^2 \cdot \left(\frac{\partial g}{\partial \bar{Y}_h} \right)^2 + \sum_{h=1}^L \frac{1}{n_h^2} \left(\frac{1}{N_h} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^3 \right) \left(\frac{\partial g}{\partial \bar{Y}_h} \right) \left(\frac{\partial^2 g}{\partial \bar{Y}_h^2} \right) \\ \quad \text{(放回抽样情况);} \\ \sum_{h=1}^L \frac{N_h - n_h}{(N_h - 1)n_h} \cdot \sigma_h^2 \left(\frac{\partial g}{\partial \bar{Y}_h} \right)^2 + \sum_{h=1}^L \frac{1}{n_h^2} \frac{(N_h - n_h)(N_h - 2n_h)}{(N_h - 1)(N_h - 2)} \\ \quad \cdot \left(\frac{1}{N_h} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^3 \right) \cdot \left(\frac{\partial g}{\partial \bar{Y}_h} \right) \left(\frac{\partial^2 g}{\partial \bar{Y}_h^2} \right) \text{(不放回抽样情况),} \end{cases} \quad (9.30)
\end{aligned}$$

其中 Y_{hi} 表示第 h 个子总体(层)中第 i 个单元.

从(9.30)可知, $V\{g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)\}$ 可以近似地表示为各子总体(层)的各阶矩的和式. 对于这些矩我们可以通过多种办法得到它们的估计. 然而这些矩前的系数主要包含了 $g(\cdot)$ 在 $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_L)$ 点的各阶偏导数, 如果用 $(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)$ 代之, 则将会引起较大偏差. Jackknife 提供了对 $g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)$ 方差的估计方法, 而不必直接涉及 $g(\cdot)$ 的有关偏导数运算. 从 $V\{g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)\}$ 近似式的各项来看, 对 $g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)$ 的 Jackknife 可以在各子总体(或层)内独立地进行. 从(9.29)来看, 这样做的结果将缩减在各子总体中所产生的 $\frac{1}{n_h}$ 阶的偏倚, 从而使整个估计量的偏倚得以缩减. 具体做法如下:

以 $\bar{y}_{h(-i)}$ 表示在第 h 个子总体(层)内舍弃第 i 个观测值后关于 \bar{Y}_h 的估计量, 记

$$g_{(k)} \triangleq g(y_1, \dots, y_{k-1}, \bar{y}_{k(-i)}, y_{k+1}, \dots, \bar{y}_L). \quad (9.31)$$

考虑到共有 L 个子总体(层), 以及(9.29)中抽样为不放回时的偏倚形式而需要添加“修正”因子, Jackknife 虚拟值定义为

$$g_{hi} = (LW_h + 1)g(\bar{y}_1, \dots, \bar{y}_L) - LW_h g_{(hi)} \\ (h = 1, 2, \dots, L; i = 1, 2, \dots, n_h). \quad (9.32)$$

其中 $W_h = \begin{cases} (n_h - 1), & \text{放回抽样情况;} \\ (n_h - 1)(1 - n_h/N_h), & \text{不放回抽样情况.} \end{cases}$

因而, 作为这些虚拟值的平均而定义的有关 $g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)$ 的 Jackknife 形式为

$$g_J(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L) = \sum_{h=1}^L \sum_{i=1}^{n_h} g_{hi} / Lm_h. \quad (9.33)$$

它几乎是 $g(\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_L)$ 的无偏估计. 现在讨论 $g(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_L)$ 的 Jackknife 方差估计. 注意到 9.30) 展示了 $V\{g(\bar{y}_1, \dots, \bar{y}_L)\}$ 的主要成份可以近似地分解为各子总体(层)的矩的线性组合, 因此定义 Jackknife 方差估计为:

$$v_1(\hat{\theta}) = \sum_{h=1}^L \frac{W_h}{n_h} \sum_{i=1}^{n_h} (g_{(hi)} - g_{(h\cdot)})^2, \quad (9.34)$$

其中 $g_{(h\cdot)} = \frac{1}{n_h} \sum_{i=1}^{n_h} g_{(hi)}.$

为了计算 $E\{v_1(\hat{\theta})\}$, 仅需研究(9.34)的和式中关于每一个 h 的期望值:

$$\frac{W_h}{n_h} E\left\{\sum_{i=1}^{n_h} (g_{(hi)} - g_{(h\cdot)})^2\right\} \\ = \frac{W_h}{n_h} E\left\{\sum_{i=1}^{n_h} (g_{(hi)} - g)^2 - n_h \left[\sum_{i=1}^{n_h} \frac{(g_{(hi)} - g)}{n_h}\right]^2\right\}. \quad (9.35)$$

关于 $g_{(hi)} - g$ ($i = 1, 2, \dots, n_h$) 在 $\bar{Y} = (Y_1, \dots, Y_L)$ 处 Taylor 展开, 经整理后得

$$(9.35) = \frac{W_h}{n_h} \left\{ \sum_{i=1}^{n_h} \left(\frac{\partial g}{\partial Y_h} \right)^2 E(\bar{y}_{h(-i)} - \bar{Y}_h)^2 \right. \\ + \sum_{i=1}^{n_h} \left(\frac{\partial g}{\partial Y_h} \right) \left(\frac{\partial^2 g}{\partial Y_h^2} \right) E(y_{h(-i)} - \bar{Y}_h)^3 \\ + \frac{1}{4} \sum_{i=1}^{n_h} \left[\left(\frac{\partial^2 g}{\partial Y_h^2} \right)^2 E(\bar{y}_{h(-i)} - \bar{Y}_h)^4 \right. \\ + \sum_{j=1}^{n_h} \left(\frac{\partial^2 g}{\partial Y_h \partial Y_j} \right)^2 E(\bar{y}_{h(-i)} - \bar{Y}_h)^2 (\bar{y}_j - \bar{Y}_j)^2 \Big] + \dots \\ \left. - n_h \left[\left(\frac{\partial g}{\partial Y_h} \right)^2 E(y_h - \bar{Y}_h)^2 \right. \right.$$

$$\begin{aligned}
& + \left(\frac{\partial g}{\partial Y_h} \right) \left(\frac{\partial^2 g}{\partial \bar{Y}_h^2} \right) E \left(\sum_{j=1}^{n_h} \frac{\bar{y}_{h(-j)} - \bar{Y}_h}{n_h} \right) \\
& \times \left(\sum_{k=1}^{n_h} \frac{y_{h(-k)} - \bar{Y}_h}{n_h} \right) + \dots \Bigg\}. \quad (9.36)
\end{aligned}$$

由于各子总体(层)的抽样相互独立,因此,经过计算近似地得到

$$\begin{aligned}
E\{v_1(\hat{\theta})\} \approx & \left\{ \begin{aligned} & \sum_{h=1}^L \frac{1}{n_h} \sigma_h^2 \left(\frac{\partial g}{\partial \bar{Y}_h} \right)^2 + \sum_{h=1}^L \frac{1}{(N_h - 1)n_h} \left(\frac{\partial g}{\partial \bar{Y}_h} \right) \left(\frac{\partial^2 g}{\partial \bar{Y}_h^2} \right) \\ & \times \left[\frac{1}{N_h} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^3 \right] \text{(放回抽样情况)}, \\ & \sum_{h=1}^L \frac{1}{n_h} \left(\frac{N_h - n_h}{N_h - 1} \right) \left(\frac{\partial g}{\partial \bar{Y}_h} \right)^2 \sigma_h^2 \\ & + \sum_{h=1}^L \frac{1}{n_h(n_h - 1)} \frac{N_h - n_h}{N_h} \left(\frac{\partial g}{\partial \bar{Y}_h} \right) \left(\frac{\partial^2 g}{\partial \bar{Y}_h^2} \right) \\ & \times \frac{(N_h - n_h + 1)(N_h - 2n_h + 2)}{(N_h - 1)(N_h - 2)} \left[\frac{1}{N_h} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^3 \right] \\ & \text{(不放回抽样情况)}. \end{aligned} \right. \quad (9.37)
\end{aligned}$$

9.3.3 弃 d -Jackknife 方差估计

Jackknife 方法可以用于统计量的方差估计,主要取决于:

1. 所构造的虚拟值有一定的散布程度.
2. 虚拟值的散布程度恰好体现了统计量本身离散程度的主要部分.

如果上述两点不成立,那末意味着 Tukey 猜想不成立,利用 Jackknife 方法进行方差估计也只能化为泡影.最著名的反例即为当 $n = k$ 时样本中位数的情况.此时不管 n 有多大,所得到的 n 个虚拟值至多只能取 3 个数值.也就是说,这些虚拟值相当“凝聚”,从而它们不能刻画样本中位数本身的离散程度.

一般来说,Jackknife 对于次序统计量的“光滑”线性组合是有效的.对于单纯的分位数则是 Jackknife 方法的一个致命点.而社会经济抽样调查经常要涉及样本分位数,例如某行业职工收入的中位数,人口抽样调查中关心的年令分位数,某产品的最小或最大寿命,人体尺寸分位数等等.根据前面的分析,要解决此类统计量的方差估计问题,看来应解决虚拟值过于“凝聚”的现象,那么最好的办法就是将每次舍弃一个观测值改成舍弃若干(设为 d , $1 < d < n$)观测值.这些舍弃后所构成的统计量将从

原来 n 个(舍弃 1 个观测值情况)增加到 $\binom{n}{d}$ 个, 有效地缓解了“凝聚”程度. 在这种想法下, 弃 d -Jackknife 方法应运而生, 它对于样本分位数的方差估计的确是一个有效的方法.

回顾本书所介绍的抽样调查的基本思想, 无非是将来自总体的某个随机样本作为该总体的一个缩影, 若以这个“缩影”作为新的总体, 重复原来的抽样程序与估计手段. 自然会有助于我们对原来抽样模型的进一步认识. 假如以 $\hat{\theta}_n$ 估计总体参数 θ , 以样本 (y_1, y_2, \dots, y_n) 为新的总体再作样本量为 $r (= n-d)$ 的不放回抽样 (这相当于从原来样本中舍弃 d 个) 得 $(y_1^*, y_2^*, \dots, y_r^*)$, 以 $\hat{\theta}_r^* = \hat{\theta}(y_1^*, y_2^*, \dots, y_r^*)$ 作为 $\hat{\theta}_n$ 的模拟, 于是这些 $\hat{\theta}_r^*$ 的离散程度恰好提供了 $\hat{\theta}_n$ 的方差的一定信息. 我们不妨以最常用的统计量——平均值加以阐述:

某有限总体的单元记作 Y_1, Y_2, \dots, Y_N , 抽样样本记作 y_1, y_2, \dots, y_n , 我们以 \bar{y} 作为 \bar{Y} 的估计. 由 § 2.2.2 知, $V(\bar{y}) = \frac{S^2}{n}(1-f)$, 其中 $S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$, 而 $f = n/N$. 因此, 要获知 $V(\bar{y})$, 我们仅需对 S^2 作出适当的估计. 利用弃 d -Jackknife 方法, 从 (y_1, y_2, \dots, y_n) 中不放回地抽取 $y_1^*, y_2^*, \dots, y_r^*$, 那末 $\bar{y}^* = \frac{1}{r} \sum_{i=1}^r y_i^*$ 是 \bar{y} 的一个模拟. 由公式(2.12), 若以 V_* 表示在弃 d -Jackknife 那样的再抽样模型下进行的方差运算, 那么

$$V_*(\bar{y}^*) = \frac{1}{r} \left(1 - \frac{r}{n}\right) s^2 = \frac{d}{nr} \cdot \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2. \quad (9.38)$$

通常 s^2 是 S^2 的无偏估计, 由(9.38)式, 我们可以用

$$\frac{nr}{d} V_*(\bar{y}^*)$$

作为 S^2 的估计. \bar{y}^* 是弃 d -Jackknife 样本的均值, 因此它的可能值有

$\binom{n}{r} = \binom{n}{d}$ 个, 从而对 $V_*(\bar{y}^*)$ 进行精确计算得

$$V_*(\bar{y}^*) = \frac{1}{\binom{n}{r}} \sum_r (\bar{y}^* - \bar{y})^2. \quad (9.39)$$

这里 \sum_r 表示对所有 $\binom{n}{r}$ 个可能的组合 $(i_1, \dots, i_r) \subset (1, 2, \dots, n)$ 求和.

因此 S^2 可以用

$$V_{Ja} \triangleq \frac{nr}{d \binom{n}{r}} \sum_r (\bar{y}^* - \bar{y})^2 \quad (9.40)$$

作为估计, 这就是 S^2 的弃 d -Jackknife 估计. 由此也就容易得到 \bar{y} 的方差估计.

现在再考虑一般的统计量, 应当指出: 弃 d Jackknife 方法也并不是对所有的统计量均可进行方差估计的. 由数理统计知识, 有些统计量 $\hat{\theta}_n$ 近似地可以表达为观测值的某种形式的均值及余项之和. 若相比之下余项相当地小, 那末该统计量几乎拥有均值的各类统计性质 (诸如相合性、渐近正态性等), 此时该统计量的方差的 n 倍即可采用如下弃 d -Jackknife 估计:

$$v_J = \frac{nr}{d \binom{n}{r}} \sum_r (\hat{\theta}_r^* - \hat{\theta}_n)^2. \quad (9.41)$$

常用的样本分位数就具有这样的性质.

当我们关心某指标的分位点时, 最经常的情况是该指标本身就是一个连续变量. 例如在人体测量中人的各部位尺寸, 比如身高的分位点, 身高本身是个连续变量; 或者指标几乎可以用一个连续分布近似地刻画, 比如关心的是中国人年龄的中位数, 由于中国人口的众多, 因此年龄的分布几乎可以看成是一个连续分布. 基于这样的认识, 我们可以将总体所有的单元 Y_1, Y_2, \dots, Y_N 看成为来自连续分布 $F(t)$ 的独立同分布观测值, 对 $F(t)$ 可以作出如下合乎常理的假设:

(*) $F(t)$ 有唯一的 p 分位点 $\xi_p (0 < p < 1)$, 其密度函数 $f(t)$ 有有界导数, 且 $f(\xi_p) > 0$.

在上述假设下, 我们用不放回抽样所得的样本 (y_1, y_2, \dots, y_n) 为基底所作的经验分布函数 $F_n(t)$ 作为 $F(t)$ 的一个近似. (y_1, y_2, \dots, y_n) 的 p 分位数可记作 $F_n^{-1}(p)$, 按照有限总体的分位数有关理论, 当 $n, N \rightarrow \infty$ 且 $N/n \rightarrow \infty$ 时,

$$\frac{\left(\frac{1}{n} - \frac{1}{N}\right)^{-1/2} (F_n^{-1}(p) - \xi_{p,N})}{\sqrt{p(1-p)/f(\xi_p)}} \xrightarrow{\mathcal{D}} N(0, 1), \quad (9.42)$$

其中 $\xi_{p,N}$ 表示 (Y_1, Y_2, \dots, Y_N) 的 p 分位数. 利用 (9.42) 式人们不难

建立 $\xi_{p,N}$ 的置信区间. 问题在于 $\left(\frac{1}{n} - \frac{1}{N}\right)^{-1/2} F_n^{-1}(p)$ 的渐近方差为 $p(1-p)/f^2(\xi_p)$, 其中 $f(\xi_p)$ 常常是未知的, 因此(9.42)式在实际应用中失去意义. 而利用弃 d -Jackknife 将出色地解决这个问题. 设 $(y_1^*, y_2^*, \dots, y_r^*) (r = n - d)$ 表示来自 (y_1, y_2, \dots, y_n) 的不放回样本, 我们以 $F_r^{*-1}(p)$ 表示 $(y_1^*, y_2^*, \dots, y_r^*)$ 的 p 分位数, 利用 Shi(1991) 的结果, 我们可以证得如下定理:

定理 9.4 假如 $F(t)$ 满足假设(*), 且令 $r = [\mu n]$, 其中 $[x]$ 表示 x 的最小整数部分, 且 $0 < \mu < 1$. 那么, 当 $n \rightarrow \infty$ 时, 有

$$\frac{n^r}{d \binom{n}{r}} \sum_r \{F_r^{*-1}(p) - F_n^{-1}(p)\}^2 \rightarrow \frac{p(1-p)}{f^2(\xi_p)} \quad \text{a.s.}, \quad (9.43)$$

将(9.42)与(9.43)相结合, 则在实际应用中就具有一定的价值.

注 1 利用弃 d -Jackknife 方法解决方差估计在理论上有其成功之处, 但在实践中产生的麻烦是公式中求和号 \sum_r 后面的项数有 $\binom{n}{r}$ 个, 当 n 比较大时, $\binom{n}{r}$ 简直是个天文数字, 我们不可能将它们全部一一列出并进行计算. 通常采用的方法是从 $\binom{n}{r}$ 个 $\hat{\theta}_r^*$ 值中随机地任取 B 个, 也就是说, 从 (y_1, y_2, \dots, y_n) 中不放回地选取 $(y_1^*, y_2^*, \dots, y_r^*)$ 以构成 $\hat{\theta}_r^*$ 的步骤重复 B 次, 得 $\hat{\theta}_{r,1}^*, \hat{\theta}_{r,2}^*, \dots, \hat{\theta}_{r,B}^*$, 我们用 $\frac{1}{B} \sum_{j=1}^B (\hat{\theta}_{r,j}^* - \hat{\theta}_n)^2$ 替代 $\frac{1}{\binom{n}{r}} \sum_r (\hat{\theta}_r^* - \hat{\theta}_n)^2$. 以上重复再抽样步骤可以在计算机上实现, 模拟次数 B 的大小直接关系到方差估计的精度, B 越大, 则估计自然越精确, 但 B 过大就会失去模拟的意义. 模拟次数 B 究竟取多大最为合适呢? 这是个尚未彻底解决的问题. 许多统计工作者从各种不同的角度要求出发进行了探讨, 据国内外的一些应用实践经验表明, 如果仅考虑方差估计, 一般地, B 大约可取在 200~1000 次之间为宜.

注 2 假如 k, m 为正整数, 且恰有 $km = n$, 若将 (y_1, y_2, \dots, y_n) 随机分为 k 组, 每组 m 个单元. 则

$$(y_1, y_2, \dots, y_n) = (y_{11}, \dots, y_{1m}; y_{21}, \dots, y_{2m}; \dots, y_{k1}, \dots, y_{km}),$$

在注 1 中所述及的模拟中, 我们取 $r = m$, $d = n\left(1 - \frac{1}{k}\right)$, 而在 $\binom{n}{r}$ 个 $\hat{\theta}_r^*$ 中

仅取这样的 k 个 $\hat{\theta}_{ri}^*$: $\hat{\theta}_{ri}^*(y_{i1}, y_{i2}, \dots, y_{im})$ ($i=1, 2, \dots, k$), 此时

$$\frac{r}{dk} \sum_{i=1}^k (\hat{\theta}_{ri}^* - \hat{\theta}_n)^2 = \frac{1}{k(k-1)} \sum_{i=1}^k (\hat{\theta}_{ri}^* - \hat{\theta}_n)^2. \quad (9.44)$$

不难看出, 这恰为 $\hat{\theta}_n$ 的随机组方差估计. 因此随机组方差估计可以看作弃 d -Jackknife 方差估计时一种特定的模拟.

9.3.4 Bootstrap 方差估计

在上一段中, 我们把从样本观测值 (y_1, y_2, \dots, y_n) 中不放回地抽取 $(y_1^*, y_2^*, \dots, y_r^*)$ 称为弃 $d(=n-r)$ Jackknife 抽样. 如果将这种再抽样的方式改为放回抽样, 则称从 (y_1, y_2, \dots, y_n) 中放回抽取的 $(y_1^*, y_2^*, \dots, y_m^*)$ 为 Bootstrap 抽样. 两者在形式上的差别主要是再抽样过程中的放回与否. 众所周知, 当 n 相当大时, 从概率论角度来看, 这种区别有点显得微不足道. 因此使人们意识到利用 Bootstrap 抽样也能对复杂样本方差估计作出贡献. 具体办法如下:

基于 Bootstrap 样本 $(y_1^*, y_2^*, \dots, y_m^*)$, 依 $\hat{\theta}_n$ 的结构构造统计量

$$\hat{\theta}_m^* = \hat{\theta}_m(y_1^*, y_2^*, \dots, y_m^*),$$

重复 Bootstrap 抽样 B 次, 相应得到 $\hat{\theta}_{m1}^*, \hat{\theta}_{m2}^*, \dots, \hat{\theta}_{mB}^*$, 于是

$$\hat{\sigma}^2 = \frac{1}{B-1} \sum_{i=1}^B (\hat{\theta}_{mi}^* - \hat{\theta}_n)^2 \quad (9.45)$$

提供了 $V(\hat{\theta}_n)$ 的估计. 通常如果将 m 改为 n (注意由于这里是放回抽样, 故 $m=n$ 是可行的, 这一点与 Jackknife 抽样有所不同.), 那么

$$v_B(\hat{\theta}_n) = \frac{1}{B-1} \sum_{i=1}^B (\hat{\theta}_{ni}^* - \hat{\theta}_n)^2 \quad (9.46)$$

提供了统计量 $\hat{\theta}_n$ 的方差估计.

Bootstrap 方法的基本思想是: 既然经验分布函数是总体分布的良好拟合, 那末来自总体分布的随机观测值的概率习性可以用经验分布函数的相应统计量的概率习性来近似地刻划, 而后者是可以通过计算机模拟甚至直接计算而得到. Bootstrap 抽样由于采用放回方式, 因此只要再抽样样本量 m 适当大 (最好 $m=n$), 对分位数这样的统计量来说, $\hat{\theta}_n$ 的 Bootstrap 模拟值 $\hat{\theta}_m^*$ (或 $\hat{\theta}_n^*$) (可能值达到 n^m 或 n^n 个) 不会发生过于“凝聚”的现象, 也就是说, 只要 m (及 n) 充分大, 分位数的 Bootstrap 方差估计将获得成功. 它不像 Jackknife 那样要求 r, d 同时充分大. 统计学家认为 Bootstrap 方法优于 Jackknife 方法的一个强有力的依据是 Bootstrap 适用于样本分位数, 而弃 1-Jackknife (即原始的 Jackknife)

不适用于分位数。但应当指出,在抽样调查中,最经常用的是从有限总体中不放回地抽取 n 个样本单元,从模拟的角度出发,人们会发现,Jackknife 的再抽样方式比起 Bootstrap 再抽样方式显得更切合实际模型。因此在实用中,如果总体是连续分布,则许多工作者偏于愿意使用 Bootstrap 法;而若遇到的是有限总体的不放回抽样,则不少工作者偏于采用 Jackknife 法。这两种再抽样方法到底哪种好?不少统计学家试图进行比较,但尚未见到全面的满意的结果。

应该强调的是:这两种方法都有其各自适用的一定局限的范围。最能说明问题的是在抽样调查中人们常关心的最大(或最小)次序统计量,它们“几乎”都不能成功(我们用“几乎”两字是指在极少数场合有成功的可能),其原因在直观上是不难理解的:用样本的最大值 $y_{(n)}$ 作为总体最大值 $Y_{(n)}$ 的估计,本身存在一个负偏差,而再抽样的最大值与样本的最大值之间又增加了一个负偏差,这样,负偏差的累积影响模型的拟合程度。而在关于观测值的光滑的统计量情况中,很少会出现这种单向的偏差累积,因此对于后者再抽样模拟常常会取得成功。

9.3.5 d, r 的选取及模拟次数 B 的确定

在弃 d -Jackknife 方法中,仅假定 r 与 d 都应随 n 增大而充分地大,究竟 r, d 之间成何种比例为宜?这是个有趣的问题。Wu(1991)从统计量分布函数的拟合,探索过这个问题。

由于有限总体单元数 N 相当大,有时我们不妨将它视作为某连续母体 $F(t)$,总体均值相当于 $F(t)$ 的中心 μ ——我们所关心的参数,记总体标准差为 σ 。通常用样本均值 \bar{y} 估计 μ ,由于总体被视作连续母体 $F(t)$,因此抽样观测值可视为独立同分布变量,此时按照 \bar{y} 分布的 Edgeworth 展开,容易得到

$$\begin{aligned} H(t) &= P \left\{ \frac{\sqrt{n}(\bar{y} - \mu)}{\sigma} \leq t \right\} \\ &= \Phi(t) + (1 - t^2)\phi(t) \frac{E(y - \mu)^3}{6\sqrt{n}\sigma^3} + o\left(\frac{1}{\sqrt{n}}\right), \end{aligned} \quad (9.47)$$

其中 $\Phi(t), \phi(t)$ 分别表示标准正态变量的分布函数及密度函数。对于弃 d -Jackknife 抽样,如果在这个模型下进行的概率运算记作 P_* 的话,那末根据不放回简单随机抽样样本均值分布的 Edgeworth 展开,可以得到(以 f 表示再抽样时的抽样比 $\frac{r}{n}$):

$$\begin{aligned}
 J(t) &= P_* \left\{ \left(\frac{nr}{d} \right)^{1/2} \frac{\bar{y}^* - \bar{y}}{\left[\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right]^{1/2}} \leq t \right\} \\
 &= \Phi(t) + (1-t^2)\phi(t) \frac{1-2f}{6\sqrt{f(1-f)}} \frac{n^{-1} \sum_{i=1}^n (y_i - \bar{y})^3}{\sqrt{n} \left[\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \right]^{3/2}} \\
 &\quad + o\left(\frac{1}{\sqrt{n}}\right). \tag{9.48}
 \end{aligned}$$

比较 $H(t)$ 与 $J(t)$, 并假定 $F(t)$ 具有足够阶矩, 显然有

$$\begin{aligned}
 n^{-1} \sum_{i=1}^n (y_i - \bar{y})^3 &\rightarrow E(y - \mu)^3, \\
 \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 &\rightarrow \sigma^2.
 \end{aligned}$$

为了使 $|H(t) - J(t)| = o\left(\frac{1}{\sqrt{n}}\right)$

成立, 当且仅当

$$\frac{1-2f}{\sqrt{f(1-f)}} \rightarrow 1 \text{ 或 } f = \frac{r}{n} \rightarrow \frac{5}{10} \sqrt{\frac{5}{10}} = 0.2764. \tag{9.49}$$

这就是说, 对于样本均值, 当每次舍弃样本单元数的 72.36% 时, Jackknife 分布拟合将达到理想程度. 一般地可适用于弃 d -Jackknife 的统计量常可以近似地表达为独立同分布变量的均值, 因此这个结论对许多统计量也有参考价值.

但是, 如果面临的问题是只考虑利用再抽样方法以解决方差估计, 那么为了分布拟合最佳选择的 r, d 未必能使方差估计达到理想的精度. 回想 Jackknife 方法本身, 之所以将每次舍弃 1 个观测值增加到舍弃 d 个, 无非是将 $\hat{\theta}$ 的可能值从 n 个增加到 $\binom{n}{r}$ 个, 从而缓解了“凝聚”程度而体现出了统计量的“离散”程度. 由此启发, 一个相当自然的想法是: 是不是这种可能值的个数越多, 则 $\hat{\theta}_n$ 本身的离散程度就会体现得更加清晰呢? 倘若这种想法对的话, 那么不妨假设 n 为偶数, 此时我们只需取 $r = d = \frac{n}{2}$ 即可达到目的. 事实上, 这种想法对某些统计量来说, 不失为是一种好的选择, 尤其是对样本中位数等那些重要的统计量, 在实际应用中证实了该想法的可行性与有效性. 具体的论述、数据的模拟以及与其他再抽样方法的比较, 我们将在 § 9.4 半样本方法中给出详尽的讨论.

在再抽样模拟过程中,究竟需要多少次计算机重复?我们在前面曾提到,用于方差估计的模拟次数 B 大约需 200~1000 次左右,但这仅仅是凭借经验与实践所提出的建议.从理论上来说, $B = \infty$ 当然是最佳选择,然而这在实践中毫无意义.倘若能保证一定的精度,那么 B 值取得越小越好.我们仅考虑 Bootstrap 方差估计的情况(在该问题上,Jackknife 与 Bootstrap 几乎无甚差异),Efron(1987)从变异系数的角度出发研究了 B 的选择.

引用 Efron 的记号,以 $\hat{\sigma}_B^2$ 表示 $\hat{\theta}$ 的 Bootstrap 方差估计,在给定抽样 $\underline{y} = (y_1, y_2, \dots, y_n)$ 的条件下,标准差的 Bootstrap 估计 $\hat{\sigma}_B$ 具有如下条件的变异系数:

$$\text{Cv}\{\hat{\sigma}_B | \underline{y}\} \approx \left[\frac{\hat{\delta} + 2}{4B} \right]^{1/2}, \quad (9.50)$$

其中 $\hat{\delta}$ 表示 $\hat{\theta}$ 的 Bootstrap 分布的峰态.在观测向量 \underline{y} 给定的情况下,当 $B \rightarrow \infty$ 时, $\text{Cv}\{\hat{\sigma}_B | \underline{y}\} \rightarrow 0$, 并且 $\hat{\sigma}_B$ 收敛于标准差的理想的 Bootstrap 估计 $\hat{\sigma}$.当然,即使在 $B \rightarrow \infty$ 时得到的 $\hat{\sigma}$ 与真正的标准差 $\sigma = \text{SD}_\theta\{\hat{\theta}\}$ 仍有所差异.令 $\text{Cv}(\hat{\sigma})$ 为 $\hat{\sigma}$ 的变异系数,那末对 \underline{y} 所有可能的实现向量取平均,就可以得到 $\hat{\sigma}_B$ 的无条件的 Cv , 近似地表示为:

$$\text{Cv}(\hat{\sigma}_B) \approx \left[\text{Cv}^2(\hat{\sigma}) + \frac{E\hat{\delta} + 2}{4B} \right]^{1/2}, \quad (9.51)$$

$\text{Cv}(\hat{\sigma})$ 有时可以在理论上计算或近似地得到.例如,若 $n = 20$, $\hat{\theta} = \bar{X}$, $X_i \stackrel{\text{iid}}{\sim} N(0, 1)$, 那么 $\text{Cv}(\hat{\sigma}) \approx (1/40)^{1/2} = 0.16$. 根据关系式(9.51),我们有可能在 Bootstrap 方差估计时适当选择 B . 以一个简单的例子说明之,假定 $E\hat{\delta} = 0$, 我们对不同数值的 B 与 $\text{Cv}(\hat{\sigma})$ 来观察 $\text{Cv}(\hat{\sigma}_B)$, 具体数据见表 9.1.

表 9.1 显示,在 $E\hat{\delta} = 0$ 假定下,当 $\text{Cv}(\hat{\sigma})$ 取 0.10 以上的值时, $B =$

表 9.1 标准差的 Bootstrap 估计 $\hat{\sigma}_B$ 的变异系数其中假定 $E\hat{\delta} = 0$

		B				
		20	50	100	200	∞
$\text{Cv } \hat{\sigma}$	0.25	0.29	0.27	0.26	0.25	0.25
	0.20	0.24	0.22	0.21	0.21	0.20
	0.15	0.21	0.18	0.17	0.16	0.15
	0.10	0.17	0.14	0.12	0.11	0.10
	0.05	0.15	0.11	0.09	0.07	0.05
	0	0.14	0.10	0.07	0.05	0

100 以上所对应的 $OV(\hat{\sigma}_B)$ 无多大改变, 因此在这种场合下, B 取 100 或 200 已经足够. 在实际操作时, 当模拟次数到达或超过某种程度时, 相应的 $\hat{\sigma}_B^2$ 数值进入“稳定”状态. 也就是说, 继续增大 B 值也不能使 $\hat{\sigma}_B^2$ 发生较大变化, 这个事实已经暗示了模拟次数 B 应该取多人为宜.

估计量的方差估计的另一个目的是获取待估参数的置信区间. 利用再抽样方法求置信区间时并不需要将估计量的方差直接计算出来, 因为根据再抽样理论, 我们可以在计算机上直接模拟出 $(\hat{\theta}_n - \theta)$ 的分布. 现记 $\hat{\theta}_n$ 的再抽样模拟为 $\hat{\theta}_i^*$ ($i = 1, 2, \dots, B$. 这里 i 表示第 i 次模拟), 那末 $(\hat{\theta}_1^* - \hat{\theta}), (\hat{\theta}_2^* - \hat{\theta}), \dots, (\hat{\theta}_B^* - \hat{\theta})$ 这 B 个模拟值构成的经验分布函数实质上就是 $(\hat{\theta}_n - \theta)$ 的分布函数 (记为 G_n) 的再抽样模拟, 将其记作 G_{nB}^* , 当 $B \rightarrow \infty$ 时, $G_{nB}^* \rightarrow G_n^*$ (G_n 的真正再抽样估计). 我们只需要将 G_{nB}^* 的分位点作为 G_n 的相应分位点的近似替代, 就可以得到 θ 的近似置信区间. 例如将 G_{nB}^* 的 2.5% 分位点 A_1 与 97.5% 分位点 A_2 视作 G_n 相应的分位点, 那么 $(\hat{\theta}_n - A_2, \hat{\theta}_n + A_1)$ 就可以近似地作为 θ 的 95% 置信区间. 这种利用计算机获得置信区间的方法为抽样调查的数据处理带来了许多方便. 问题在于这里的模拟次数 B 又如何确定呢? 它的解决与前面所述的方差估计的 B 的选择有所区别. 不少学者对此曾有所探讨. Shi, Wu 与 Chen (1990) 建立了有限总体分位点过程的 Bahadur 表示式, 从而得到了关于 B 的如下关系式:

$$(\log \log B/B)^{1/2} = c d_n \quad \text{对某些常数 } c > 0, \quad (9.52)$$

其中

$$d_n = \sup_{-\infty < x < \infty} |G_n^*(x) - G_n(x)| \quad (9.53)$$

表示 G_n 与它的再抽样估计 G_n^* 之间的距离. 关系式 (9.52) 提出了模拟次数 B 应与原始样本大小 n 以及分布的再抽样估计精度有密切的关系. 假如 d_n 较大, 由 (9.52) 式可看出可以选小一些的 B , 直观告诉我们: 当分布的再抽样估计的精度比较粗糙时, 大量地做计算机模拟无法弥补这个缺陷. 相反地, 如果 d_n 比较小, 即再抽样估计的精度令人满意的话, 我们应当增大 B , 否则, 由于计算机模拟次数少, 而带来的偏倚将使再抽样方法提供的精度蒙受损失, 这是十分可惜的事情.

无论是方差估计或是置信区间, 要对它们提供一个确切的 B 值是几乎不可能的. 我们只能指出模拟次数的趋向或有关它的阶数.

Jackknife 与 Bootstrap 方法对于样本均值具有令人满意的效果. 例如弃 1-Jackknife 作用于 y , 则估计量形式保持不变, 而 Jackknife 方差

估计与通常采用公式也一致。因此,为说明再抽样用于估计统计量方差的效应,常常采用非线性统计量作为例子。具体方法是从一个已知的总体(连续母体或有限总体)中随机放回(或不放回)地抽取 n 个样本,构造所要求的统计量,此时统计量的方差是可以计算得到的。对于该统计量实施再抽样技巧可以得到它的方差估计,然后与已知的方差比较。使用偏倚及均方误差,以评估再抽样方法的效应。

Jackknife 与 Bootstrap 方差估计对于样本分位数具有较理想的效果,我们所作的有关中位数的模拟结果将在下节半样本估计中一起列出。

§ 9.4 半样本方法

9.4.1 基本思想与方法

在随机组方法中,最简单的做法是将样本分为两组,不妨假设 n 为偶数: $y_1, \dots, y_{\frac{n}{2}}, y_{\frac{n}{2}+1}, \dots, y_n$, 则关于 $\hat{\theta}_n$ 的随机组方差估计为:

$$\frac{1}{2}\{[\hat{\theta}_{\frac{n}{2}}(y_1, \dots, y_{\frac{n}{2}}) - \hat{\theta}_n]^2 + [\hat{\theta}_{\frac{n}{2}}(y_{\frac{n}{2}+1}, \dots, y_n) - \hat{\theta}_n]^2\}.$$

将 n 个样本单元分为两组,共有 $\binom{n}{\frac{n}{2}}$ 种可能,若任取一种代入上述公式,

作为方差估计量,就存在由于偶然性而影响精度的可能。为克服由这种

偶然性带来的麻烦,最好的方法是将所有 $\binom{n}{\frac{n}{2}}$ 种可能统统代入公式,然后

对所得到的结果加以平均,利用公式则可表示为:

$$\left(\frac{n}{2}\right)^{-1} \sum_{\frac{n}{2}} \{\hat{\theta}_{\frac{n}{2}}(y_{i_1}, \dots, y_{i_{\frac{n}{2}}}) - \hat{\theta}_n\}^2, \quad (9.54)$$

即为 $\hat{\theta}_n$ 的弃 $\frac{n}{2}$ -Jackknife 方差估计量。前面的分析表明,直观上它有可能使方差估计达到较佳的效果。注意到在随机组方法及 Jackknife 方

法的公式中,可以将 $\hat{\theta}_n$ 代之以全体 $\hat{\theta}_{\frac{n}{2}}$ 的平均,而使方差估计改变甚微,

于是(9.54)式成为:

$$\left(\frac{n}{2}\right)^{-1} \sum_{\frac{n}{2}} \left\{ \hat{\theta}_{\frac{n}{2}}(y_{i_1}, \dots, y_{i_{\frac{n}{2}}}) - \left(\frac{n}{2}\right)^{-1} \sum_{\frac{n}{2}} \hat{\theta}_{\frac{n}{2}}(y_{j_1}, \dots, y_{j_{\frac{n}{2}}}) \right\}^2$$

$$= \left(\frac{n}{2} \right)^{-2} \Sigma^* \frac{1}{2} \{ \hat{\theta}_{\frac{n}{2}}(y_{i_1}, \dots, y_{i_{\frac{n}{2}}}) - \hat{\theta}_{\frac{n}{2}}(y_{j_1}, \dots, y_{j_{\frac{n}{2}}}) \}^2, \quad (9.55)$$

其中 Σ^* 表示对所有 $((i_1, \dots, i_{\frac{n}{2}}), (j_1, \dots, j_{\frac{n}{2}}))$ 这样的配对求和, 而 $(i_1, \dots, i_{\frac{n}{2}}), (j_1, \dots, j_{\frac{n}{2}})$ 均为 $(1, 2, \dots, n)$ 中的容量为 $\frac{n}{2}$ 的子集. 公式(9.55)可以用一个大家相当熟悉的简单事实加以阐述:

设 x_1, \dots, x_n 为来自某总体的样本(通过放回或不放回抽样), 设该总体的方差为 σ^2 , 它的无偏估计 $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, s^2 可以改写成

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n(n-1)} \sum_{i,j} \frac{1}{2} (x_i - x_j)^2. \quad (9.56)$$

这种表示式明确地告诉我们: 随机观测值互相之间差异的平方平均值实际上在一定程度上提供了这些观测值来自总体的方差的信息, 而(9.55)与(9.56)式在形式上的一致性也明确地说明了(9.55)的确为我们提供了 $\hat{\theta}_{\frac{n}{2}}$ 方差的信息. 然而用(9.55)式右边的公式进行运算相当繁复, 而且 Σ^* 内包含着那么多项中出现的某些配对 $(i_1, \dots, i_{\frac{n}{2}})$ 与 $(j_1, \dots, j_{\frac{n}{2}})$ 之间共有元素过多的现象, 直观告诉我们: 这种现象的过多出现有可能影响关于 $\hat{\theta}_{\frac{n}{2}}$ 的方差估计的精确性. 顺乎自然的想法是: 为了减少运算量, 我们要设法减少 Σ^* 内的项数, 而为了提高精度, 则希望在 Σ^* 内舍弃的项应是那些共有元素较多的配对, 从这个角度出发, 仅留下那些没有共有元素的配对, 用以进行估计计算也许是最理想的方法. 这就引出了关于 $\hat{\theta}_{\frac{n}{2}}$ 方差的所谓半样本估计:

$$\hat{V}_{\text{HS}}(\hat{\theta}_{\frac{n}{2}}) = \frac{1}{2 \binom{n}{2}} \sum_{\frac{n}{2}} (\hat{\theta}_{\frac{n}{2}} - \hat{\theta}_{\frac{n}{2}}^c)^2. \quad (9.57)$$

在(9.57)式中用 $\hat{\theta}_{\frac{n}{2}}, \hat{\theta}_{\frac{n}{2}}^c$ 分别表示依赖于各一半的样本而构造的统计量. 鉴于我们所关心的是统计量方差的 n 倍, 即 $nV(\hat{\theta}_{\frac{n}{2}}) \triangleq \sigma^2 + R$, 其中 R 为无穷小量, 因此其主体部分 σ^2 的半样本估计为:

$$\hat{\sigma}_{\text{HS}}^2 = \frac{n}{4 \binom{n}{2}} \sum_{\frac{n}{2}} (\hat{\theta}_{\frac{n}{2}} - \hat{\theta}_{\frac{n}{2}}^c)^2, \quad (9.58)$$

即(9.57)式乘上 $\frac{n}{2}$. 若 n 为奇数, 当然无法用(9.58)式. 这时, 我们取 $r = (n+1)/2$, $d = (n-1)/2$. 将(9.54)稍作推广并仍称之为半样本估计:

$$\hat{\sigma}_{\text{HS}}^2 = \frac{(n-1)(n+1)}{4n \binom{n}{\frac{n+1}{2}}} \sum_r (\hat{\theta}_r - \hat{\theta}'_d)^2. \quad (9.59)$$

我们用 $\hat{\theta}_r$ 与 $\hat{\theta}'_d$ 表示这两个统计量不依赖于共有样本. 由于 n 相当大, (9.59)与(9.58)式几乎没有差别, 因此为简便起见, 我们只讨论 n 为偶数的情况.

9.4.2 半样本方差估计性质

本段主要考虑非线性统计量半样本方差估计的偏性. 我们总是假设总体元素个数 N 充分大(其实这就是我们进行抽样调查而不进行普查的重要理由之一). 这样的抽样方法是否为放回的, 差异不大. 因此为了使计算方便, 且使问题趋于简单起见, 在本段中仅考虑 (y_1, y_2, \dots, y_n) 为独立同分布随机观测值. 有如下简单的结论:

定理 9.5 当 n 为偶数时,

$$E\hat{\sigma}_{\text{HS}}^2 = \frac{n}{2} V(\hat{\theta}_{\frac{n}{2}}). \quad (9.60)$$

证明 利用 (y_1, y_2, \dots, y_n) 的独立同分布性, 易知 $\hat{\theta}_{\frac{n}{2}}$ 与 $\hat{\theta}_{\frac{n}{2}}^c$ 互为独立的, 故

$$\begin{aligned} & E \left[\frac{n}{4 \binom{n}{\frac{n}{2}}} \sum_{\frac{n}{2}} (\hat{\theta}_{\frac{n}{2}} - \hat{\theta}_{\frac{n}{2}}^c)^2 \right] \\ &= \frac{n}{4} \cdot \frac{1}{\binom{n}{\frac{n}{2}}} \sum_{\frac{n}{2}} \left[E(\hat{\theta}_{\frac{n}{2}} - E\hat{\theta}_{\frac{n}{2}})^2 + E(\hat{\theta}_{\frac{n}{2}}^c - E\hat{\theta}_{\frac{n}{2}}^c)^2 \right. \\ &\quad \left. - 2E(\hat{\theta}_{\frac{n}{2}} - E\hat{\theta}_{\frac{n}{2}})(\hat{\theta}_{\frac{n}{2}}^c - E\hat{\theta}_{\frac{n}{2}}^c) \right] \\ &= \frac{n}{2} E(\hat{\theta}_{\frac{n}{2}} - E\hat{\theta}_{\frac{n}{2}})^2 = \frac{n}{2} V(\hat{\theta}_{\frac{n}{2}}). \blacksquare \end{aligned}$$

这个定理的意义并不仅仅在于半样本估计量是 $\frac{n}{2} V(\hat{\theta}_{\frac{n}{2}})$ 的无偏估

计. Efron与Stein(1978)以及Bhargava(1983)都曾研究过 Jackknife 方差估计的偏性. 他们发现在一般的情况, Jackknife 方差估计具有正偏倚. 因此关于估计本身以及由此估计所提供的置信区间也趋于保守. 半样本方差估计在偏性问题上的确可能优于其他再抽样方法. 下面我们 用 Monte Carlo 的结果进行一些比较.

所介绍的几种再抽样方差估计方法对于样本均值几乎具有相同的效 果, 对于观测值的光滑函数的模拟结果也几乎是相同的. Shao 与 Shi (1989)对非光滑的中位数作了比较, 指定总体所拟合的分布分别为正 态、Cauchy 指数分布. 利用计算机模拟计算样本中位数的方差估计的偏 倚与均方误差平方根($\sqrt{\text{MSE}}$), 具体结果归纳如表 9.2 所示.

表 9.2 样本中位数方差(σ^2)的再抽样估计的偏倚与 $\sqrt{\text{MSE}}$

估计量	分布状态					
	正 态		Cauchy		指 数	
	$\sigma^2=6.28$		$\sigma^2=4.93$		$\sigma^2=6.25$	
	偏倚	$\sqrt{\text{MSE}}$	偏倚	$\sqrt{\text{MSE}}$	偏倚	$\sqrt{\text{MSE}}$
半样本 ($B=100$)	0.35	3.42	0.6	3.75	0.00	4.16
Jackknife, $d=n/2, B=100$	0.57	4.00	1.69	4.63	1.05	5.12
Jackknife, $d=n/2, B=256$	0.56	3.91	1.43	4.03	1.01	4.91
Bootstrap $B=100$	1.00	4.35	2.10	5.07	1.47	5.48
Bootstrap $B=256$	0.93	4.14	1.84	4.43	1.47	5.39
Jackknife, $d=n/4, B=256$	1.37	6.14				

注: 表中正态分布为 $N(2.5, 2^2)$, Cauchy 分布具中位数 2.5 及形状参数 $\sqrt{2}$, 指数分 布具 2.5, 中位数 1.73.

表 9.2 显示; 无论从偏倚还是从 $\sqrt{\text{MSE}}$ 的角度而论, 半样本方差估 计优于 Jackknife 与 Bootstrap 方差估计. 如果只局限于 Jackknife 方 法的话, 一般弃 $\frac{n}{2}$ 优于 d 的其他选择. 对其他统计量的模拟也显示了类 似的结论. 不过应当指出的是: 半样本方差估计的这种优势似乎对于中 位数或一些分位数的模拟结果尤为显著.

9.4.3 平衡半样本估计

一、总体均值分层抽样的估计

假如需要估计总体均值 \bar{Y} , 共有 L 层, 每层的单元数设为 $N_h (h=1, 2, \dots, L)$, $N = \sum_{h=1}^L N_h$. 从每一层中随机放回地抽取两个样本, 所获得的

$2L$ 个样本单元提供了 \bar{Y} 的一个分层估计量:

$$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h,$$

其中 $W_h = N_h/N$, $\bar{y}_h = (y_{h1} + y_{h2})/2$, 鉴于各层之间的抽样是独立进行的, 因此 $V(\bar{y}_{st})$ 的估计为

$$v(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 (y_{h1} - y_{h2})^2 / 4. \quad (9.61)$$

从每层的两个抽样中选择一个样本, 则构成所谓半样本 $(y_{1i_1}, y_{2i_1}, \dots, y_{Li_1}) (i_1, i_2, \dots, i_L \text{ 取 } 1 \text{ 或 } 2)$, 这样的可能数共有 2^L 种, 相应于第 α 组半样本 ($\alpha = 1, 2, \dots, 2^L$) 的 Y 估计量为

$$\bar{y}_{st, \alpha} = \sum_{h=1}^L W_h (\delta_{h1\alpha} y_{h1} + \delta_{h2\alpha} y_{h2}), \quad (9.62)$$

其中

$$\delta_{h1\alpha} = \begin{cases} 1, & y_{h1} \text{ 被选入第 } \alpha \text{ 组半样本;} \\ 0, & \text{其他.} \end{cases}$$

$$\delta_{h2\alpha} = 1 - \delta_{h1\alpha}.$$

容易验证

$$\begin{aligned} \sum_{\alpha=1}^{2^L} \bar{y}_{st, \alpha} / 2^L &= \sum_{\alpha=1}^{2^L} \sum_{h=1}^L W_h (\delta_{h1\alpha} y_{h1} + \delta_{h2\alpha} y_{h2}) / 2^L \\ &= \sum_{h=1}^L (y_{h1} + y_{h2}) 2^{L-1} / 2^L = \bar{y}_{st}. \end{aligned} \quad (9.63)$$

运用半样本方法的基本思想, 我们可以利用这些 $\bar{y}_{st, \alpha}$ 估计 \bar{y}_{st} 的方差. 先引进记号:

$$\begin{aligned} \delta_h^{(\alpha)} &= 2\delta_{h1\alpha} - 1 \\ &= \begin{cases} 1, & \text{若 } y_{h1} \text{ 选入第 } \alpha \text{ 组半样本;} \\ -1, & \text{若 } y_{h2} \text{ 选入第 } \alpha \text{ 组半样本,} \end{cases} \end{aligned}$$

于是

$$\bar{y}_{st, \alpha} - \bar{y}_{st} = \sum_{h=1}^L W_h \delta_h^{(\alpha)} (y_{h1} - y_{h2}) / 2, \quad (9.64)$$

$$\begin{aligned} (\bar{y}_{st, \alpha} - \bar{y}_{st})^2 &= \sum_{h=1}^L W_h^2 (y_{h1} - y_{h2})^2 / 4 \\ &+ \sum_{h < h'} W_h W_{h'} \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} (y_{h1} - y_{h2})(y_{h'1} - y_{h'2}) / 2. \end{aligned} \quad (9.65)$$

注意: 统计量 $(\bar{y}_{st, \alpha} - \bar{y}_{st})^2$ 包含一个主项与一个交叉项, 显然主项即为 $v(\bar{y}_{st})$, 而由于各层抽样的独立性, 交叉项的期望自然为 0. 这个事实提供了一个简单的想法, 2^L 个统计量 $(\bar{y}_{st, \alpha} - \bar{y}_{st})^2$ 的平均是 $V(\bar{y}_{st})$ 的一个无偏估计.

定理 9.6

$$E\left\{\sum_{\alpha=1}^{2^L} (\bar{y}_{st,\alpha} - \bar{y}_{st})^2 / 2^L\right\} = V(\bar{y}_{st}). \quad (9.66)$$

然而, 若层数 L 相当多的话, 估计量 $\sum_{\alpha=1}^{2^L} (\bar{y}_{st,\alpha} - \bar{y}_{st})^2 / 2^L$ 中所包含的计算量会使得实际操作发生困难, 再抽样方差估计的模拟方法使我们可以选择 k 项加以平均:

$$v_k(\bar{y}_{st}) \triangleq \sum_{\alpha=1}^k (y_{st,\alpha} - \bar{y}_{st})^2 / k. \quad (9.67)$$

显然这也是 $V(\bar{y}_{st})$ 的无偏估计. 对于固定的 k , 求和式中项的选择无非有两种可能供参考: 一为随机地从 2^L 个半样本中独立不放回地抽取 k 个; 二为特殊的选择以满足某种需要. 由 (9.65) 式

$$\begin{aligned} V\{v_k(\bar{y}_{st})\} &= V\{v(\bar{y}_{st})\} \\ &+ \sum_{h < h'} \frac{2^L}{k} \frac{k}{2^L - 1} W_h^2 W_{h'}^2 V(y_{h1} - y_{h2}) V(y_{h'1} - y_{h'2}) / 4. \end{aligned} \quad (9.68)$$

显然, 一般来说 $v_k(\bar{y}_{st})$ 比起 $v(\bar{y}_{st})$ 的精度差一些. 若我们能选取 k 个特殊的半样本, 以使 $v_k(\bar{y}_{st}) = v(\bar{y}_{st})$, 那末 $V\{v_k(\bar{y}_{st})\} = V\{v(\bar{y}_{st})\}$. 为了达到此目的, 由 (9.65) 式知, 对这 k 组半样本, 仅需满足

$$\sum_{\alpha=1}^k \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} = 0 \quad (9.69)$$

对一切 $h < h' = 1, 2, \dots, L$ 成立. Plackett 与 Burman (1946) 构造了 $k \times k$ (k 为 4 的倍数) 正交矩阵, 其列满足该条件. 例如 $k=8$ 时, 我们在表 9.3 中提供了 5、6、7、8 层情况的半样本中满足条件 (9.69) 的子集, 以这种方式确定所需要的 k 个半样本, 自然导致等式 $v_k(\bar{y}_{st}) = v(\bar{y}_{st})$. 也就是说, k 个半样本完全包含了 2^L 组半样本中所提供的有关 $V(\bar{y}_{st})$ 的所有信息, 而 $v_k(\bar{y}_{st})$ 中有关层的交叉分量则被“省略”了. McCarthy (1966) 称这样的 k 组半样本为平衡半样本.

平衡半样本给出一个理想的结果: $v_k(\bar{y}_{st}) = v(\bar{y}_{st})$, 同时它也导致了另外一个令人鼓舞的性质: 只要我们选取的 k 组半样本满足条件

$$\sum_{\alpha=1}^k \delta_h^{(\alpha)} = 0 \quad (9.70)$$

对 $h = 1, 2, \dots, L$ 成立, 此时 $\frac{1}{k} \sum_{\alpha=1}^k \bar{y}_{st,\alpha} = \bar{y}_{st}$ 成立. 当 $k > L$ 时, Plackett 和 Burman 方法提供了这方面的保证, 我们可以选择 k 以使 (9.70) 成立.

表 9.3 关于 5、6、7 或 8 层的平衡半样本重复的确定

半样本	层							
	2	3	4	5	6	7	8	
$\delta_k^{(1)}$	-1	1	-	+1	-1	+1	+1	-1
$\delta_k^{(2)}$	+1	+1	1	1	+1	-1	+1	-1
$\delta_k^{(3)}$	+1	+1	+1	-1	-1	+1	-1	-1
$\delta_k^{(4)}$	1	+1	+1	-1	-1	-1	+1	-1
$\delta_k^{(5)}$	+1	1	+1	+1	+1	1	1	-1
$\delta_k^{(6)}$	-1	+1	1	-1	+1	+1	-1	1
$\delta_k^{(7)}$	-1	-1	+1	-1	+1	+1	+1	-1
$\delta_k^{(8)}$	1	1	1	-1	1	1	-1	-1

注: 在 (α, h) 格中 +1 指 $(h, 1)$ 单元在第 α 组半样本中, 在 (α, h) 格中 -1 指 $(h, 2)$ 单元在第 α 组半样本中。这样的表实际上是一个 Hadamard 矩阵。

但当 $k=L$ 时, 由表 9.3 可见最后一列全为 -1, 即最后一层中某一元素在 k 组半样本内全都出现, 因此不满足条件 (9.70), 条件 (9.69) 与 (9.70) 同时成立时, 称这样的半样本组选择为完全正交平衡。

二、分层抽样时的一般估计量

在前述抽样情况下, 如果参数 θ 的估计量为 $\hat{\theta}$, 那么基于某半样本的估计量不妨记作 $\hat{\theta}_\alpha$, 基于该组半样本的余集也存在相应的半样本估计量, 记作 $\hat{\theta}_\alpha^c$ 。于是, 基于 k 个平衡半样本的关于 $V(\hat{\theta})$ 的估计可以为

$$v_k(\hat{\theta}) = \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta})^2 / k, \quad (9.71)$$

$$v_k^c(\hat{\theta}) = \sum_{\alpha=1}^k (\hat{\theta}_\alpha^c - \hat{\theta})^2 / k. \quad (9.72)$$

由于每层内两个单元的编号是对称的, 因此 (9.71) 所依据的 k 组半样本为平衡的话, 那么 (9.72) 中所依据的 k 组半样本当然是平衡的, 结合 v_k 与 v_k^c , 不难得到一个新的方差估计:

$$\hat{v}_k(\hat{\theta}) = [v_k(\hat{\theta}) + v_k^c(\hat{\theta})] / 2. \quad (9.73)$$

另外还可以运用上节提到的样本之差的平方和提供了母体方差信息的基本思想, 得到

$$v_k^*(\hat{\theta}) = \sum_{\alpha=1}^k (\hat{\theta}_\alpha - \hat{\theta}_\alpha^c)^2 / 4k. \quad (9.74)$$

对于 $\hat{\theta}$ 为观测值的线性函数, 容易验证这四个估计量其实是等同的。但当 $\hat{\theta}$ 不是线性形式时, 一般地, 它们并不相等, 对于 $v_k^*(\hat{\theta})$, 则有

$$\begin{aligned}
 E v_k^*(\hat{\theta}) &= \frac{1}{4k} \sum_{\alpha=1}^k [V(\hat{\theta}_\alpha) + V(\hat{\theta}_\alpha^c) - 2\text{Cov}(\hat{\theta}_\alpha, \hat{\theta}_\alpha^c)] \\
 &= \frac{1}{2} V(\hat{\theta}_\alpha), \\
 &\text{或} = V\left\{\frac{\hat{\theta}_\alpha + \hat{\theta}_\alpha^c}{2}\right\} \triangleq V(\hat{\theta}_\alpha), \quad (9.75)
 \end{aligned}$$

即 $v_k^*(\hat{\theta})$ 是 $V(\hat{\theta}_\alpha)$ 的无偏估计. 一般地, 当 L 较大时, 我们认为 $V(\hat{\theta}_\alpha)$ 相当接近于 $V(\hat{\theta})$, 因此通常 $v_k^*(\hat{\theta})$ 作为 $V(\hat{\theta})$ 的一个估计. 而另外三个表达式 $v_k(\hat{\theta})$ 、 $v_k^c(\hat{\theta})$ 以及 $\bar{v}_k(\hat{\theta})$ 从形式上可知常作为 $\text{MSE}(\hat{\theta})$ 的估计.

三、各层抽样为简单无放回情况

我们仍假定从 L 分层各抽 2 个单元构成样本以估计总体均值 \bar{Y} . 只不过这一次各层中 2 个样本单元的抽样为随机不放回的形式. 那么得 \bar{Y} 的估计量:

$$y_{st} = \sum_{h=1}^L W_h y_h = \sum_{h=1}^L W_h (y_{h1} + y_{h2})/2, \quad (9.76)$$

且

$$\begin{aligned}
 V(y_{st}) &= \sum_{h=1}^L W_h^2 \left(1 - \frac{2}{N_h}\right) S_h^2/2 \\
 &= \sum_{h=1}^L W_h^2 \left(1 - \frac{2}{N_h}\right) (y_{h1} - y_{h2})^2/4. \quad (9.77)
 \end{aligned}$$

以 (9.77) 比较 (9.61), 易见原来的权 W_h 将用 $W_h^* = W_h \sqrt{1 - 2/N_h}$ 代替. 如果仍沿用记号 $\delta_{h1\alpha}$ 及 $\delta_{h2\alpha}$, 我们可以定义 $V(\bar{y}_{st})$ 的半样本估计:

$$\bar{y}_{st,\alpha}^* = \bar{y}_{st} + \sum_{h=1}^L W_h^* (\delta_{h1\alpha} y_{h1} + \delta_{h2\alpha} y_{h2} - \bar{y}_h), \quad (9.78)$$

$$v_k^{**}(\bar{y}_{st}) = \frac{1}{k} \sum_{\alpha=1}^k (\bar{y}_{st,\alpha}^* - \bar{y}_{st})^2. \quad (9.79)$$

v_k^{**} 的表达式也涉及到 k 组半样本的适当选择. 沿用每层有放回抽样的记号, 可以验证, 如果 k 组半样本的选择满足条件

$$\sum \delta_h^{(\alpha)} \delta_{h'}^{(\alpha)} = 0 \quad (h \neq h')$$

及

$$\sum \delta_h^{(\alpha)} = 0.$$

此时 v_k^{**} 具有下述特性:

$$(1) \quad v_k^{**}(\bar{y}_{st}) = v(\bar{y}_{st}), \quad (9.80)$$

$$(2) \quad \frac{1}{k} \sum_{\alpha=1}^k \bar{y}_{st,\alpha}^* = \bar{y}_{st}. \quad (9.81)$$

对于一般的估计量, 例如 \bar{y}_{st} 的函数 $g(\bar{y}_{st})$ (用来估计参数 $g(\theta)$), 可以得到 $g(\bar{y}_{st})$ 的方差的半样本估计:

$$v_h^{**}\{g(\bar{y}_{st})\} = \frac{1}{k} \sum_{\alpha=1}^k [g(\bar{y}_{st,\alpha}^*) - g(\bar{y}_{st})]^2. \quad (9.82)$$

显然, 当 $g(\cdot)$ 是线性函数时, 对平衡半样本组的适当选择可以使 $v_h^{**}\{g(\bar{y}_{st})\}$ 也满足 (9.80) 与 (9.81) 式. 但是, 当 $g(\cdot)$ 为非线性函数时, 通常, 无论怎样选取 k 组半样本, 都很难使 (9.80) 与 (9.81) 成立. 然而假如 $g(\cdot)$ 具有良好的函数性质时, 利用 Taylor 展开的方法, $g(\bar{y}_{st})$ 的主要习性常在很大程度上依赖于其展开式的线性部分, 因此, 只要 k 组半样本的选取满足条件 (9.69) 与 (9.70), 那么 (9.80) 与 (9.81) 式对于估计量 $g(\bar{y}_{st})$ 近似地成立.

9.4.4 每层多于两个样本单元情况

在上述 L 层内, 各层均抽两个样本的情况毕竟不多, 一般地, 我们假设每层抽样数 $n_h \geq 2$. 本节讨论的一个主题是至少有一层 $n_h > 2$ 严格成立. 为讨论方便起见, 不妨假定 n_h 均为偶数. 这样, 我们可以将 n_h 个单元随机地划分为两组, 记作 $y_{h,1} = (y_{h11}, y_{h12}, \dots, y_{h1\frac{n_h}{2}})$, $y_{h,2} = (y_{h21}, y_{h22}, \dots, y_{h2\frac{n_h}{2}})$.

这种划分共有 $\binom{n_h}{\frac{n_h}{2}}$ 种可能. 现假定估计总体均值 \bar{Y} , 我们以

$\bar{y}_{h,i}$ 表示 h 层内第 i ($i=1, 2$) 分组的平均值, 显然 $\bar{y}_h = (\bar{y}_{h,1} + \bar{y}_{h,2})/2$. 于是 \bar{Y} 的估计为

$$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h.$$

但 \bar{Y} 的半样本估计有 $\binom{n_1}{\frac{n_1}{2}} \cdot \binom{n_2}{\frac{n_2}{2}} \cdots \binom{n_L}{\frac{n_L}{2}}$ 种可能. 对于每层确定的分组

$$\bar{y}_{st,\alpha}^* = \begin{cases} \sum_{h=1}^L W_h (\delta_{h1\alpha} \bar{y}_{h,1} + \delta_{h2\alpha} \bar{y}_{h,2}) & (\text{每层样本为独立放回抽取}); \\ \bar{y}_{st} + \sum_{h=1}^L W_h^* (\delta_{h1\alpha} \bar{y}_{h,1} + \delta_{h2\alpha} \bar{y}_{h,2} - \bar{y}_h) & (\text{每层样本为不放回抽取}). \end{cases} \quad (9.83)$$

这里 $W_h, W_h^*, \delta_{h1\alpha}, \delta_{h2\alpha}$ 的意义同前. 相应的方差半样本估计为

$$v_h^{**}(\bar{y}_{st}) = \frac{1}{k} \sum_{\alpha=1}^k (\bar{y}_{st,\alpha}^* - \bar{y}_{st})^2. \quad (9.84)$$

显然, 当各层的分组均确定时, 半样本的点估计与方差估计与每层仅抽 2 个样本时具有类似形式. 关键在于每层分组不同而引起大量不同的 $y_{st,\alpha}^*$.

如何具体操作对 \bar{y}_{st} 作出半样本方差估计, 可以有各种不同的考虑:

1. 当 L 很大时, 可以用随机的方式在各层确定半样本组. 一旦确定了, 则采用各层两个样本的处理方式.

2. 当 L 比较小时, 特别是 $L=1$ 或者 $L=2$ 时, 则考虑所有可能的半样本组合.

但在实际抽样中, 我们遇到的估计量常常表现为各层估计量的线性组合, 由于各层之间抽样的独立性, 因此仅需各别进行每层中估计量的方差估计, 即可得到整个估计量的方差估计. 而在各层中, 我们自然采用 9.4.2 段中所介绍的半样本方差估计的方法.

9.4.5 部分平衡半样本估计

在一个复杂分层抽样方案中, 假如层数 L 相当大, 例如 $L=80$, 即使平衡半样本方法可以使得计算所需要的半样本项数既少又有效, 但由于 $k \geq L$ 的要求, 而使得这种方法仍显得费时费钱. 随之提出的问题是: 能否设计一组 k 个部分平衡半样本, 由此得到的方差估计, 比来自 k 个独立半样本所产生的方差估计有较理想的精度. 本段就介绍这种方法如下:

假定有 L 层 (L 相当大), 且在平衡半样本中采用 k 组, 由于 $k < L$, 因此所谓“平衡”, 只能部分地达到. 现为了叙述方便起见, 不妨假设 L 可以被 k 整除, 令 $L/k = G$, 于是我们将 L 层分为 G 群. 为使问题叙述清楚, 考虑 $L=4$ (假定这是个很大的数), 那么共有 $2^4 = 16$ 个可能半样本. 按照平衡半样本方法, 必须有 $k \geq 4$. 因此我们可以像表 9.3 那样利用 Hadamard 矩阵构造 4 组平衡半样本. 现在我们仅取 $k=2$, 则 4 层分为 $L/k=2$ 群. 对于包含第一、二层的第一群, 采用 2 阶 Hadamard 矩阵构造 2 个正交列, 而在包含第三、四层的第二群中重复第一群的方法. 具体设计如表 9.4.

表 9.4

半样本	层 (k)			
	1	2	3	4
$\delta_h^{(1)}$	+1	+1	+1	+1
$\delta_h^{(2)}$	+1	-1	+1	-1

此时,显然有

$$\begin{aligned} v_2(y_{st}) = & \frac{1}{4} \sum_{k=1}^4 W_k (y_{k1} - y_{k2})^2 \\ & + \frac{1}{2} \{ W_1 W_3 (y_{11} - y_{12})(y_{31} - y_{32}) \\ & + W_2 W_4 (y_{21} - y_{22})(y_{41} - y_{42}) \}. \end{aligned} \quad (9.85)$$

我们看到由于设计的特点,除了两群中有相同向量的层之间的交叉项保留外,其余的交叉项已经全部抵消.因此采用该设计所得到的半样本方差估计比较平衡半样本估计多了若干交叉项,但却减少了不少计算量,而它比随机独立地选取 k 组半样本的方法又减少了许多交叉项,起到了平衡半样本方法的某些效应.利用这种设计方法得到半样本方差估计的方法称为部分平衡半样本方法.对于一般的 L, k (只要 L 是 k 的整数倍),在 L/k 个群的每一群中利用 Hadamard 矩阵构造 k 列正交向量,每群的构造方法完全一样,则得到 k 组半样本构成如下的方差估计:

$$\begin{aligned} & \sum_{\alpha=1}^k (y_{st,\alpha} - \bar{y}_{st})^2 / k \\ & = \frac{1}{4} \sum_{h=1}^L W_h^2 (y_{h1} - y_{h2})^2 + \frac{1}{2} \sum_{h,j} W_h W_j (y_{h1} - y_{h2})(y_{j1} - y_{j2}). \end{aligned} \quad (9.86)$$

其中第二个和号是对符合以下条件的所有 (h, j) 求的:

$$h < j,$$

h 来自某一群 k 层,而 j 来自另一群 k 层,

h 与 j 表示两群中相应的 Hadamard 矩阵的相同列的两个层.

第二个和号中共包含 $k \cdot \frac{L}{k} \left(\frac{L}{k} - 1 \right) / 2 = \frac{L(L-k)}{2k}$ 项.

显然由于各层之间抽样为独立的,部分平衡半样本方差估计仍为无偏估计.

§ 9.5 Taylor 级数法

在实际抽样调查中,除了总体均值、总体总和等参数可以用观测值的线性函数形式作估计之外,还常常运用一些非线性估计量,诸如比估计量、相关系数、回归系数等等.通常这些非线性估计量的方差没有精确表示式,当然也就谈不上简单无偏估计.

倘若这些非线性估计量中的某一类可以用样本观测值的线性函数作为近似, 那么再运用已有的关于线性估计量的方差估计的方法, 至少可以得到一个虽然有偏但确是相合的方差估计. 这种线性近似的方法主要依赖于 Taylor 展开或者二项展开的有效性. 需要强调的是: Taylor 展开本身并不能估计方差, 它仅仅提供估计量的一个线性逼近, 然后再利用前面所介绍的方法以得到近似的方差估计.

9.5.1 估计量方差的线性近似估计

考虑一个给定的有限总体 N , 令 $Y = (Y_1, \dots, Y_p)'$ 表示总体参数的 p 维向量. 在 Taylor 级数展开的大多数应用中, 这 p 个参数 $Y_i (i=1, 2, \dots, p)$ 通常是 p 个不同的调查指标的全体总和或均值. 因此基于 n 个样本单元的关于 Y_i 的估计量一般采用标准的估计量 \hat{Y}_i , 通常 \hat{Y}_i 是 Y_i 的无偏估计, 有时即使是有偏但相合性较好的估计量. 于是 Y 的估计量当然采用样本向量 $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_p)'$.

假如我们感兴趣的参数并不仅是 Y , 而是 Y 的函数形式 $\theta = g(Y)$. 那末自然采用估计量 $\hat{\theta} = g(\hat{Y})$. 现在面临的问题是:

- (1) 寻找 $\hat{\theta}$ 的设计方差的近似表达式;
- (2) 对 $\hat{\theta}$ 的方差建立一个适当的估计量.

正如前面所述, 我们总是考虑 \hat{Y} 是 Y 的良好估计, 甚至为无偏估计. 而对 $\hat{\theta} = g(\hat{Y})$ 采用 Taylor 级数的另一个前提则为函数 $g(\cdot)$ 具有相当光滑的性质. 例如假定在包含 Y 与 \hat{Y} 的某个开集 (这样的开集的存在由于参数 \hat{Y} 的未知, 一般较难核实, 好在我们处理的函数 $g(\cdot)$ 通常定义域即为此所要求的开集) 内具有二阶连续偏导数, 于是, 由数学分析中常规的 Taylor 展开得到:

$$\begin{aligned} \hat{\theta} - \theta &= \sum_{j=1}^p \frac{\partial g(Y)}{\partial Y_j} (\hat{Y}_j - Y_j) \\ &\quad + \frac{1}{2} \sum_{j=1}^p \sum_{k=1}^p \frac{\partial^2 g(\bar{Y})}{\partial Y_j \partial Y_k} (\hat{Y}_j - Y_j)(\hat{Y}_k - Y_k), \end{aligned} \quad (9.87)$$

其中 \bar{Y} 位于 \hat{Y} 与 Y 之间, 注意到 (9.87) 式与 (9.28) 式的相似, 因此从理论上讲, 我们可以采用 Jackknife 方差估计方法得到 $V(\hat{\theta})$ 的近似估计, 而且避免了关于函数 $g(\cdot)$ 的偏导数运算. 但是, 如果 $g(\cdot)$ 的偏导数容易计算的话, 那么 Taylor 级数展开仍不失为试图估计方差的有效手段.

在有限总体中, 通常认为 (9.87) 式右端第二项相对于 $\hat{\theta} - \theta$ 是个“不

重要”的分量, 于是 $\hat{\theta}$ 的均方误差近似为:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E\{g(\hat{P}) - g(Y)\}^2 \\ &\doteq V\left\{\sum_{j=1}^p \frac{\partial g(Y)}{\partial Y_j} (\hat{P}_j - Y_j)\right\} \\ &= \sum_{j=1}^p \sum_{i=1}^p \frac{\partial g(Y)}{\partial Y_j} \frac{\partial g(Y)}{\partial Y_i} \text{Cov}(\hat{P}_j, \hat{P}_i) \triangleq d \hat{\Sigma} d', \end{aligned} \quad (9.88)$$

其中 $\hat{\Sigma}$ 为 \hat{P} 的协方差矩阵, $d, d_j = \frac{\partial g(Y)}{\partial Y_j} (j=1, \dots, p)$. 注意到 $\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + \text{Bias}^2(\hat{\theta})$, $V(\hat{\theta})$ 与 $\text{MSE}(\hat{\theta})$ 为同阶, 而 $\text{Bias}^2(\hat{\theta})$ 为高阶无穷小, 因此就一阶近似来说, $V(\hat{\theta})$ 与 $\text{MSE}(\hat{\theta})$ 的估计是一致的. 仅需在(9.88)式中, 把 $\hat{\Sigma}$ 代之以样本估计 $\hat{\Sigma}$, 把 d 代之以 \hat{d} , 其中 $\hat{d}_j = \frac{\partial g(\hat{P})}{\partial Y_j}$. 因此所得估计为 $v(\hat{\theta}) = \hat{d} \hat{\Sigma} \hat{d}'$. 当然利用进一步 Taylor 展开, 还可以继续得到二阶或更高阶的近似. 这对于比较若干种方差(或均方误差)估计方法的优劣, 显然是有意义的. 但是, 如果我们的目的仅仅在于估计方差, 那么一般大型的复杂抽样调查显示一阶近似常常可以产生比较满意的结果. 需要注意的一点是: 倘若总体发生严重偏倚, 则依赖近似不可能使人们满意.

(9.88)式可以不费什么力气而推广到多元的情况. 倘若我们感兴趣的 q 维参数向量可以表示为

$$G(Y) = (g_1(Y), g_2(Y), \dots, g_q(Y))',$$

那么, 相应的估计量自然采用形式

$$G(\hat{P}) = (g_1(\hat{P}), g_2(\hat{P}), \dots, g_q(\hat{P}))'.$$

此时 $G(\hat{P})$ 的均方误差矩阵与交叉项近似为

$$E\{[G(\hat{P}) - G(Y)][G(\hat{P}) - G(Y)]'\} \doteq D \hat{\Sigma} D'. \quad (9.89)$$

其中 D 为 $q \times p$ 阶矩阵, 其一般元为.

$$d_{ij} = \frac{\partial g_i(Y)}{\partial Y_j}.$$

这样, 我们又得到了 $G(\hat{P})$ 的方差-协方差估计为

$$V(G(\hat{P})) = \hat{D} \hat{\Sigma} \hat{D}', \quad \hat{d}_{ij} = \frac{\partial g_i(\hat{P})}{\partial Y_j}.$$

利用 Taylor 级数展开进行方差估计的有效性存在着可能令人们怀疑之处, 主要表现在以下两点:

(1) 用以得到(9.88)式的 Taylor 展开是否收敛? 如果不收敛, 那么

(9.88)式作为 $\text{MSE}(\hat{\theta})$ 的近似表示式显然是不合适的.

(2) 如果 Taylor 展开收敛, 则收敛速度又是一个令人关心的问题. 因为收敛速度直接影响到近似方差估计的精度.

这两个问题在连续总体模型时容易处理, 因为在该模型下有可能建立 Taylor 展开式余项的阶数, 并且可以发现比起展开式的线性项来, 余项具有较高阶的无穷小量. 这样, 在近似的过程中, 可以略去余项, 可是对于有限总体模型来说, 如果不对该模型作出一定的假设, 就比较难有前述的结果.

例如, 设 \bar{y} 与 \bar{x} 表示基于大小为 n 的不放回抽样的样本均值, 它们之间的比值 $\hat{R} = \bar{y}/\bar{x}$ 用以估计总体均值比 $R = \bar{Y}/\bar{X}$. 现令 $\delta_y = (\bar{y} - \hat{Y})/\hat{Y}$, $\delta_x = (\bar{x} - \bar{X})/\bar{X}$. 则可记 $\hat{R} = R(1 + \delta_y)(1 + \delta_x)^{-1}$, 在 $\delta_x = 0$ 处展开 \hat{R} 为 Taylor 级数的形式:

$$\begin{aligned}\hat{R} &= R(1 + \delta_y)(1 - \delta_x + \delta_x^2 - \delta_x^3 + \delta_x^4 - \cdots) \\ &= R(1 + \delta_y - \delta_x - \delta_y\delta_x + \delta_x^2 - \cdots).\end{aligned}$$

此级数当且仅当 $|\delta_x| < 1$ 时收敛. 因此对所有 $\binom{N}{n}$ 个可能样本来说, 当且仅当 $|\delta_x| < 1$ 时, (9.88)式关于 $\text{MSE}(\hat{R})$ 的近似公式才会成立.

针对上述情况, Koop 曾构造了一个违反该收敛条件的简单例子(可参见 Wolter(1985)):

例 9.2 某总体 $N = 20$, 各单元分别取值为: 5, 1, 3, 6, 7, 8, 1, 3, 10, 11, 16, 4, 2, 11, 6, 6, 7, 1, 5, 13. 此时 $\bar{x} = 6.3$. 选定某容量大小为 4 的样本组, 有 $\bar{x} = (11 + 16 + 11 + 13)/4 = 12.75$, 于是 $\delta_x = (\bar{x} - \bar{X})/\bar{X} = 6.45/6.3 > 1$. 若取 $n = 2$ 或 3 时, 也存在某些样本组, 使得 $|\delta_x| > 1$ 成立. 但是当样本组的容量大小增加到 5 时, 条件 $|\delta_x| < 1$ 恒成立. Koop 称 $n = 5$ 为临界样本容量.

上面这个例子启示我们, 当样本量增加时, 那些由“极端”观测值引起麻烦的可能性会相应地减少, 从而增大 Taylor 展开式收敛的可能. 由经验显示, 如果我们采用相当有效的调查方案, 并且使样本容量充分大, 那么一阶 Taylor 级数展开常常提供可靠的近似. 在有限总体的抽样调查中, 有关非线性估计量广泛采用一阶近似.

9.5.2 应用 Taylor 级数于特殊的估计量

Hansen、Hurwitz 与 Madow(1953)讨论了一种特殊的情况, 所感兴

趣的参数具有如下形式:

$$\theta = g(Y) = \frac{Y_1 \cdot Y_2 \cdots Y_m}{Y_{m+1} \cdot Y_{m+2} \cdots Y_p}, \quad (9.90)$$

其中 $1 \leq m \leq p$, 最简单的例子为比 $\theta = Y_1/Y_2$, 相应的估计量采用

$$\hat{\theta} = \frac{\hat{Y}_1 \cdot \hat{Y}_2 \cdots \hat{Y}_m}{\hat{Y}_{m+1} \cdot \hat{Y}_{m+2} \cdots \hat{Y}_p}, \quad (9.91)$$

其中 \hat{Y}_i 是 Y_i 的标准估计, 假如它是无偏估计. 利用 Taylor 级数展开, 可以得到 $\text{MSE}(\hat{\theta})$ 的一阶近似:

$$\begin{aligned} \text{MSE}(\hat{\theta}) \approx & \theta^2 \{ [\sigma_{11}/Y_1^2 + \cdots + \sigma_{mm}/Y_m^2] \\ & + [\sigma_{m+1,m+1}/Y_{m+1}^2 + \cdots + \sigma_{pp}/Y_p^2] \\ & + 2[\sigma_{12}/(Y_1 Y_2) + \sigma_{13}/(Y_1 Y_3) \\ & + \cdots + \sigma_{m-1,m}/(Y_{m-1} Y_m)] \\ & + 2[\sigma_{m+1,m+2}/(Y_{m+1} Y_{m+2}) + \cdots \\ & + \sigma_{p-1,p}/(Y_{p-1} Y_p)] \\ & - 2[\sigma_{1,m+1}/(Y_1 Y_{m+1}) + \sigma_{1,m+2}/(Y_1 Y_{m+2}) \\ & + \cdots + \sigma_{m,p}/(Y_m Y_p)] \}, \end{aligned} \quad (9.92)$$

其中 $\sigma_{ij} = \text{Cov}(\hat{Y}_i, \hat{Y}_j)$ 为矩阵 Σ 的基本元. 如果对于 σ_{ij} ($i, j = 1, 2, \dots, p$) 存在恰当的估计 $\hat{\sigma}_{ij}$, 那么 $\text{MSE}(\hat{\theta})$ 可以用下式进行估计:

$$\begin{aligned} v(\hat{\theta}) = & \hat{\theta}^2 \{ [\hat{\sigma}_{11}/\hat{Y}_1^2 + \cdots + \hat{\sigma}_{mm}/\hat{Y}_m^2] \\ & + [\hat{\sigma}_{m+1,m+1}/\hat{Y}_{m+1}^2 + \cdots + \hat{\sigma}_{pp}/\hat{Y}_p^2] + 2[\hat{\sigma}_{12}/(\hat{Y}_1 \hat{Y}_2) \\ & + \hat{\sigma}_{13}/(\hat{Y}_1 \hat{Y}_3) + \cdots + \hat{\sigma}_{m-1,m}/(\hat{Y}_{m-1} \hat{Y}_m)] \\ & + 2[\hat{\sigma}_{m+1,m+2}/(\hat{Y}_{m+1} \hat{Y}_{m+2}) + \cdots \\ & + \hat{\sigma}_{p-1,p}/(\hat{Y}_{p-1} \hat{Y}_p)] - 2[\hat{\sigma}_{1,m+1}/(\hat{Y}_1 \hat{Y}_{m+1}) \\ & + \hat{\sigma}_{1,m+2}/(\hat{Y}_1 \hat{Y}_{m+2}) + \cdots + \hat{\sigma}_{m,p}/(\hat{Y}_m \hat{Y}_p)] \}. \end{aligned} \quad (9.93)$$

(9.92) 式及 (9.93) 式是容易记的, 对于所有的一般项 (i, j) 来说, 当 $i=j$ 时, 该项有一个相应的方差 σ_{ii} (或方差估计 $\hat{\sigma}_{ii}$) 除以相应指标的平方 Y_i^2 (或 \hat{Y}_i^2), 该项前面的系数为 +1, 而当 $i \neq j$ 时, 则有一个相应于变量 \hat{Y}_i 与 \hat{Y}_j 的协方差 (或协方差的估计) 除以两个相应指标 (或它们的估计) 的乘积, 该项前面的系数取 ± 2 , 当 i, j 两个相应指标同在分子或同在分母时取 +2, 否则取 -2.

考虑参数 (9.90) 式的最简单的比以及比估计的情况, 此时 $\theta = R = \bar{Y}/\bar{X}$, 则 $\hat{R} = \hat{y}/\hat{x}$, \hat{x} 、 \hat{y} 均为不放回抽样估计. 利用一阶 Taylor 近似, $\text{MSE}(\hat{R})$ 的估计可取为

$$\begin{aligned}
 v(\hat{R}) &= \hat{R}^2 (\sigma_y^2/\bar{y}^2 + \sigma_x^2/\bar{x}^2 - 2\sigma_{xy}/\bar{x}\bar{y}) \\
 &= \frac{\bar{y}^2}{x^2} \left\{ \frac{1-f}{n\bar{y}^2} \cdot s_y^2 + \frac{1-f}{nx^2} \cdot s_x^2 - \frac{2(1-f)}{nxy} s_{xy} \right\} \\
 &= \frac{1-f}{nx^2} (s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}). \quad (9.94)
 \end{aligned}$$

注意第4章比估计的方差估计中的公式(4.59), 那里的 $v_1(\hat{R})$ 与(9.94)式中的 $v(\hat{R})$ 是一致的.

在第11章案例分析中给出了 Taylor 级数法的一个很好的实例. 在 § 11.5 的《1987 年中国儿童情况抽样调查》中考虑了某两个指标的比估计问题. 对于比估计量的方差估计, 利用 Taylor 级数可以由(9.94)式解决. 该案例采用了分层二阶不等概率整群抽样方法, 从而使 Taylor 级数方差估计式中 $\sigma_{xy} = \text{Cov}(\hat{X}, \hat{Y})$ 成为关键. 在该案例的处理中主要采取了先将待估的 $\text{Cov}(\hat{X}, \hat{Y})$ 近似地表达成两个容易估计的参数的线性组合, 从而最终解决了问题.

9.5.3 鞍点逼近方法

我们在本章曾数次提到过, 在抽样调查实践中, 有时关心的指标本身是连续变量, 例如人体的身高、体重等, 有时由于有限总体单元个数相多, 而相应地将关心指标近似视作为连续变量. 在这种情况下, 我们可以借助于近几十年来越来越受到国际统计学界瞩目的鞍点逼近的方法, 以得到统计量的分布密度(或分布函数), 从而获得统计量的方差估计, 尤其是可以获得统计量分布的分位数, 提供了参数估计精度的一类刻划. 下面简略地介绍这种鞍点逼近方法:

假设 X_1, X_2, \dots, X_n 为独立变量且具有共同的分布 F , 我们试图获得 $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$ (或等价地 $\sum_{j=1}^n X_j$) 的密度近似, 令 $K(\lambda) = \log E(e^{\lambda X})$ 表示变量 X 的累积量母函数(cumulant generating function), 则在 t 点 \bar{X} 的密度可以表示为(本段中 i 表示纯虚数单位):

$$f(t) = \frac{n}{2\pi i} \int_C \exp[n\{K(\lambda) - \lambda t\}] d\lambda, \quad (9.95)$$

其中 C 是 $K(\lambda)$ 收敛域内的复围道. 利用鞍点逼近方法(Reid(1988)对鞍点逼近与统计推断有一个出色的综述)可以得到 \bar{X} 在 t 点的密度的鞍点展开为:

$$g(t) = \left\{ \frac{n}{2\pi K''(\lambda_t)} \right\}^{\frac{1}{2}} \exp[n\{K(\lambda_t) - \lambda_t t\}](1 + R_n), \quad (9.96)$$

其中 λ_t 称作鞍点, 是鞍点方程 $K'(\lambda) = t$ 的根; K' 、 K'' 分别表示 $K(\lambda)$ 的一阶、二阶导数; 而余项 R_n 一般具有 n^{-1} 次幂的展开.

在相当广泛的条件下, 统计学家证明了鞍点方程存在唯一实根. 特别地, 包括了随机变量 X_j 可以服从离散分布的情况.

(9.96)式的重要性有两点: 首先其误差阶为 n^{-1} 而不是中心极限定理中正态逼近的 $n^{-\frac{1}{2}}$ 阶; 其次, 误差是一致的, 从而在分布的尾部逼近即使 n 比较小仍然相当精确.

对于 \bar{X} 的累积分布函数 G 相应的近似式由 Daniels (1987) 给出. 若 F 为连续分布的话, 则有

$$G(t) \approx \Phi(w_t) + \phi(w_t)(w_t^{-1} - z_t^{-1}), \quad (9.97)$$

其中

$$w_t = [2n\{\lambda_t t - K(\lambda_t)\}]^{\frac{1}{2}} \text{sgn}(\lambda_t),$$

$$z_t = \lambda_t \{nK''(\lambda_t)\}^{\frac{1}{2}}.$$

$\text{sgn}(\cdot)$ 为符号函数, λ_t 仍为鞍点方程的根, $\Phi(\cdot)$ 、 $\phi(\cdot)$ 分别表示标准正态分布函数与密度函数.

利用(9.97)式可以得到 G 的各个分位点. 倘若我们考虑的是 $\bar{X} \sim \mu$ ($\mu = E\bar{X}$) 的分布的话, 那么实际上得到了参数 μ 的有关置信区间. 但是(9.97)式的缺点在于需要知道 F , 否则, 就无法解出鞍点 λ_t 来. 如果 F 未知, 我们可以用 X_1, X_2, \dots, X_n 的经验分布函数 \hat{F}_n 来代替, 此时累积量母函数 $K(\lambda)$ 可以估计为

$$\hat{K}(\lambda) = \log \left\{ n^{-1} \sum_{j=1}^n \exp(\lambda x_j) \right\}. \quad (9.98)$$

因而所求鞍点为下述方程的唯一解:

$$t = \frac{\sum_{j=1}^n x_j \exp(\lambda x_j)}{\sum_{j=1}^n \exp(\lambda x_j)}. \quad (9.99)$$

上述方程在 $t \leq \min(x_j)$ 与 $t \geq \max(x_j)$ 时无解, 这在实际中是相当清楚的.

在理论上, 以经验分布函数 \hat{F}_n 代替 F , (9.97)式的误差阶不再是一致的, 但是利用鞍点逼近方法仍然有精确的效果. Davison 和 Hinkley (1983, 对于 $(\bar{X} - \mu)$ 给出了一个数值例子如下

给定一组 $n=10$ 的样本:

9.6 10.4 13.0 15.0 16.6,
17.2 17.3 21.8 24.0 33.8.

$(\bar{X} - \mu)$ 的精确分位点只能通过 Bootstrap 模拟得到, 即从这 10 个数据中放回地抽取 10 个 Bootstrap 数据构成 (\bar{X}^*, \bar{x}) , 为了使之尽可能精确, 取模拟次数为 50000 次. 具体结果见下表:

表 9.5 $\bar{X} - \mu$ 的再抽样分位点的近似

概率	精确值	鞍点近似	正态近似
0.0001	-6.34	-6.31	-8.46
0.0005	-5.74	-5.78	-7.48
0.001	-5.55	-5.52	-7.03
0.005	-4.81	-4.81	-5.86
0.01	-4.42	-4.43	-5.29
0.05	-3.34	-3.33	-3.74
0.10	-2.69	-2.69	-2.91
0.20	-1.86	-1.86	-1.91
0.80	1.80	1.80	1.91
0.90	2.87	2.85	2.91
0.95	3.73	3.75	3.74
0.99	5.47	5.48	5.29
0.995	6.12	6.12	5.86
0.999	7.52	7.46	7.03
0.9995	8.19	7.99	7.48
0.9999	9.33	9.12	8.46

†. 精确值一栏由 50000 次 bootstrap 模拟而得.

从上表可以看出: 鞍点逼近得到的分位点比常用的正态逼近精确得多, 尤其是对于分布两端的分位点更显示出鞍点逼近的优越性. 另外我们还可以看到, 利用鞍点逼近公式可以取得大量与 Bootstrap 模拟几乎同样的效果.

鉴于在样本量 n 比较小的情况, 鞍点逼近可以提供 $\bar{X} - \mu$ 的分布接近尾部分位点的相当精确的估计, 因此, 在有限总体的抽样调查中, 人们已经开始注意到这种方法的实用价值.

第 10 章

非抽样误差及相关问题

在第 9 章中介绍的关于复杂样本的方差估计, 以及前面几章给出的关于一些简单估计量方差估计的公式, 都基于这样一种信念: 无论在调查中采用哪一种抽样方案, 我们所得到的每一个观测值 y 均是正确无误的, 前面所谈及的误差, 是由于企图用局部(n 个抽样单元的数据)去推断总体过程中必然会发生的差异, 为避免局部“替代”整体时发生“极端”的情况, 我们采用随机的手法获取 n 个样本, 讨论的正是由于随机抽样过程中所产生的随机误差, 即抽样误差。

倘若整个抽样调查过程比较简单, 而且拥有相当高级的计算器具, 并且我们的调查人员极端认真负责…。这一切保证了上述假定的可能性。然而, 一旦进入抽样调查实践, 人们常常发现这些假定是不尽人意的, 在复杂的抽样调查中尤其如此。在实践中, 除了抽样误差外, 可能产生误差的来源常见如下:

一、无回答现象

对某些选定的样本在调查过程中发生计算遗漏。最突出的表现在所关心的指标涉及到人, 而被调查者要么找不到, 要么拒绝回答。

二、调查误差

由于工具或人为的一些因素而造成观测值 y 与真正的 Y 有偏误。

三、资料数据整理过程中所产生的误差

例如调查数据的登录及计算机录入过程中发生的错误。

这些误差的可能存在, 使得我们不能依照前面几章所讲述的方法计算误差及置信限。因而我们面临的新问题是如何减少这些非抽样误差, 在某些非抽样误差的确存在的情况下, 又如何去有效地计算误差与置信限。

§ 10.1 无回答及其影响

10.1.1 无回答的类型

无回答(non-response)的类型粗略地归结为:

一、遗漏

由于样本抽取的随机性, 存在某些已经确定要去调查的单位发生找不到的现象, 或者由于客观存在的困难, 诸如交通不便, 气候恶劣等而使得无法找到被调查者。

二、不在家

被调查者恰好不在家, 通常有两种情况可处理, 一种是该家庭中其他人可以作出回答, 这种处理比较容易, 另一种调查比较注重对象选择的随机性, 选择到谁, 就调查谁, 于是由于“不在家”而引起了无回答现象。

三、不能回答

包括某些被调查人对所调查的问题缺少有关资料或者不愿意提供。

1993年我们在对一些企事业单位调查有关职工收入, 住房等指标时发生了不少第一和第二两类无回答现象。有些单位主管负责人外出形成了“不在家”, 而其他人所提供资料常常不准确; 有些单位缺乏其中某些资料, 甚至有些单位明确表示不愿意提供这方面的数据, 经过若干工作后, 一般单位还较愿意配合。但也存在一些如下第四类“无回答”现象。

四、坚决拒绝调查

由此产生的偏差一般难于消除。

10.1.2 无回答的影响

任何一种调查总可以将总体分为两个部分: 一部分是一旦抽到就可以得到回答并进行计量的单元, 设该部分总数为 N_1 ; 另一部分由一旦入样会产生“无回答”的单元所组成, 设其总数为 N_2 ($N_1 + N_2 = N$)。这两个部分的划分当然与所关心的指标、所查找的单位以及采用调查的手法密切相关。例如人口抽样调查倘若只关心性别、出生年月等指标, 一般“无回答”部分所占比例相当小。若询问的是年收入与年支出分配这样的问题, 则 N_2 可能相对大。又若调查采用多次访问再加上调查员的工作细致周到, 比较只采用一次访问的调查, 前者的“无回答”部分就可能比后者小得多。

现设 $W_1 = N_1/N$, $W_2 = N_2/N$, 假如采用简单随机抽样来估计总体均值, 此时我们手中仅有第一部分所得样本数据, 于是得到偏倚为:

$$E(\hat{y}_1) - Y = \bar{Y}_1 - Y = \bar{Y}_1 - (W_1\bar{Y}_1 + W_2\bar{Y}_2) = W_2(\bar{Y}_1 - \bar{Y}_2). \quad (10.1)$$

上式中 W_2 与 \bar{Y}_1 均可以估得, 然而 \bar{Y}_2 是样本所无法提供的, 因此无法获知偏倚的大小.

假如所关心的指标是个连续变量, 其可能取值范围相当大, 于是 Y_2 的取值范围有可能也相当大, 加上无回答部分所占的比例 W_2 相当大时, 我们无法从样本获知偏倚, 更无法确定 \bar{Y} 的置信限. 如果硬要获得有关 \bar{Y} 的置信限, 唯一可行的方法是对偏倚作一些猜测, 当然这样的猜测常常无法证实其正确性, 从而所得的“置信限”缺乏一定的依据, 缺乏相当的精确度. 可见, 无回答现象的存在对于抽样推断影响很大.

假如所关心的是连续型指标 θ , 如同总体均值一样, 具有形式:

$$\theta = W_1\theta_1 + W_2\theta_2. \quad (10.2)$$

如果 θ 是个取值范围有限的连续变量, 一般地 θ_2 也当如此. 对于总体的第一部分指标 θ_1 , 具有样本估计 $\hat{\theta}_1$, 利用经典统计及抽样理论, 则可得到 θ_1 的(在一定置信度下)上、下置信限—— θ_{1U} 及 θ_{1L} . 假如 θ_2 如同 θ 一样具有取值的上、下限: A_U 与 A_L , 那么在理论上可得到 θ 的置信上、下限:

$$W_1\theta_{1L} + W_2A_L < \theta < W_1\theta_{1U} + W_2A_U. \quad (10.3)$$

假定 W_2 已知, 那么上述区间可以粗略地作为 θ 的置信区间. 由于 θ_2 采用了其可能取值的两端, 因此这样得到的置信区间显然是保守的, 也就是说, θ 落在该区间的可能性大于所给定的置信度. 这一点在理论上也不难证明. 而且可以清晰地看到, 该区间的长度与 W_2 的大小很有关系. W_2 愈小, 区间长度就愈短, 置信度就愈接近于设计方案所要求的. 因此, 在整个抽样调查过程中, 有时值得花费一部分资金作多次访问等进一步工作, 以便减少无回答的比例.

假如 W_2 未知, 而调查的回答只有两种可能: “是”或“否”. 那么我们可以得到一个更粗略一些的置信区间, 只要在计算 θ 的上下限时将无回答者的反应全部认可为回答“是”(或“否”). 当然, 这种看来是自然的估计, 会引起置信区间的长度增大.

10.1.3 多次访问及其模型

为了缩小无回答的影响, 减小无回答的数量, 我们有必要采取一些措

施,例如采用多次访问的方法。当然多次访问对于那些“坚决拒绝回答”者来说很难奏效,但对于“不在家”、“不能回答”等无回答类型,是有一定作用的。问题在于如何确定多次访问的次数,以便既在经济上承担得起,又能减少无回答数量。

确定多次访问的次数是个较困难的事情,它涉及到调查的方案、访问的相对费用以及花费的时间等各种因素。

假如调查的内容是被访问单位(或家庭)中其他任何人都能回答的,那么第一次访问的成功率显然较高。故访问的次数常常不必规定太多。但当要去调查被随机抽中的人时,一般第一次访问的成功率将比前种情况要小得多,但是由于调查员能乘第一次扑空时可能会了解到被调查者何时在家等信息,第二、三次访问的成功率将显著比前种情况增加,因此在这种方案下,一般规定的访问的次数要略多一些。

关于调查的相对费用,主要关心的是到某次访问结束,按全部完成的调查通摊计算每一完成的调查的平均费用。利用费用的计算从而估计能在资金方面承担的访问次数,有时需要利用历史的或经验的资料进行测算。

另外,时间方面的考虑对确定多次访问的次数也是重要的,众所周知,多次访问必定会延迟取得最后结果的时间。这就是需要根据调查的时间方面的要求加以考虑。

Deming(1953)建立了一个多次访问效果的数学模型;

根据找到被调查者的概率将总体划分为 r 组,引进一些记号;

w_{ij} = 在 i 次访问中找到第 j 组的一个被调查对象的概率(不妨假设 $w_{ij} > 0$); p_j = 总体中属于第 j 组的比例;

μ_j = 第 j 组某指标的均值; σ_j^2 = 第 j 组某指标的方差。

以 \bar{y}_{ij} 表示在总共 i 次访问中所找到的第 j 组中被调查对象有关指标的均值,假定

$$E(\bar{y}_{ij}) = \mu_j, \quad (10.4)$$

于是,该指标的总体均值为:

$$\bar{\mu} = \sum_{j=1}^r p_j \mu_j. \quad (10.5)$$

对于确定要访问的样本,经过第 i 次访问之后可以划分为 $(r+1)$ 组。样本属于第 1 组并被调查者;样本属于第 2 组并被调查者, ..., 依次类推,直到样本属于第 r 组并被调查者,至于样本中的第 $(r+1)$ 组则由第 i 次访

问后尚未被调查者组成。粗略地, 我们可以认为这 $(r+1)$ 组中的人数 $(m_1, m_2, \dots, m_{r+1})$ 服从多项分布, 而 $m_1 + m_2 + \dots + m_{r+1} = n_0$ 为调查方案确定要调查的样本总数。因此,

$$n_i = m_1 + m_2 + \dots + m_r \quad (10.6)$$

表示在 i 次访问过程中被调查过的总人数, 那么该随机变量显然服从成功概率为 $w_{i1}p_1 + w_{i2}p_2 + \dots + w_{ir}p_r$, 试验次数为 n_0 的二项分布。它的期望应为

$$E(n_i) = n_0 \sum_{j=1}^r w_{ij}p_j. \quad (10.7)$$

而当 n_i 固定时, (m_1, m_2, \dots, m_r) 又服从各别成功概率为 $w_{ij}p_j / \sum_j w_{ij}p_j$ 的多项分布, 因而

$$E(m_j | n_i) = \frac{n_i w_{ij} p_j}{\sum_j w_{ij} p_j}. \quad (10.8)$$

我们以 \bar{y}_i 表示 i 次访问之后得到的样本均值, 则

$$E(\bar{y}_i | n_i) = E\left(\frac{\sum m_j \bar{y}_{ij}}{n_i}\right) = \frac{\sum n_i w_{ij} p_j \mu_j}{n_i \sum w_{ij} p_j} = \frac{\sum w_{ij} p_j \mu_j}{\sum w_{ij} p_j} \triangleq \bar{\mu}_i. \quad (10.9)$$

这个条件期望的结果并不依赖于条件变量 n_i 的取值, 表示了 \bar{y}_i 的无条件期望当然也为 $\bar{\mu}_i$ 。于是得估计量 \bar{y} 的偏差是 $(\bar{\mu}_i - \bar{\mu})$ 。同样, 我们可以求得给定 n_i 时, y_i 的条件方差:

$$V(y_i | n_i) = \frac{\sum_j w_{ij} p_j [\sigma_j^2 + (\mu_j - \bar{\mu}_i)^2]}{n_i \sum_j w_{ij} p_j}. \quad (10.10)$$

注意到

$$\begin{aligned} \frac{1}{n_i} &= \frac{1}{(En_i) \left(1 + \frac{n_i - En_i}{En_i}\right)} \\ &= \frac{1}{En_i} \left\{ 1 - \frac{n_i - En_i}{En_i} + \left(\frac{n_i - En_i}{En_i}\right)^2 - \dots \right\}. \end{aligned} \quad (10.11)$$

如果忽略掉二阶以上无穷小, 可以近似地得到 y_i 的无条件方差, 即规定作 i 次访问情况下样本均值的方差:

$$V(\bar{y}_i | i) \approx \frac{\sum_j w_{ij} p_j [\sigma_j^2 + (\mu_j - \bar{\mu}_i)^2]}{n_0 \left(\sum_j w_{ij} p_j\right)^2} \quad (10.12)$$

因而得到在按规定作 i 次访问之后所得估计量 \bar{y}_i 的均方误差:

$$\text{MSE}(\bar{y}_i | i) = V(\bar{y}_i | i) + (\bar{\mu}_i - \bar{\mu})^2. \quad (10.13)$$

按照(10.13)式,我们可以大致地估算访问次数 i 的取值,以使 $\text{MSE}(\bar{y}_i)$ 达到调查设计的要求,其间,必须要结合考虑进行 i 次访问所花费用. 第 k 次($k=1, 2, \dots, i$)访问中完成的调查(不包含第 k 次以前的调查)的期望值容易计算为 $\sum (w_{kj} - w_{k-1,j})p_j$. 假定 O_k 为第 k 次访问中每一个完成调查的平均费用,那么,进行 i 次访问的总费用期望值为 $n_0 O(i)$, 其中

$$O(i) = O_1 \sum w_{1j} p_j + O_2 \sum (w_{2j} - w_{1j}) p_j + \dots + O_i \sum (w_{ij} - w_{i-1,j}) p_j. \quad (10.14)$$

将费用及均方误差综合考虑多次访问的问题,一般是费用支出额为固定的情况,所得 $\text{MSE}(\bar{y})$ 数值自然随访问次数不同而改变. 如果我们从历史与经验中积累一些有关费用及相关的 w_{ij} 、 p_j 以及相对偏差等资料,那么有可能比较 $\text{MSE}(\bar{y})$. 假如只进行一次访问,给定的费用可以支付 n_0 个抽样,随着我们规定访问次数的增多,利用有关资料可以分别求出 n_1, n_2, \dots, n_i 的期望值,从而利用公式分别求得 $V(\bar{y})$ 及 $\text{MSE}(\bar{y})$.

10.1.4 校正无回答误差的方法

无回答现象的存在,对抽样统计推断产生一定的影响. 这种影响随着无回答部分所占比例的增加而扩大. 因此在实际抽样调查中,应当采取一些措施,以校正由于无回答而产生的误差.

(一)对第一次访问后的“无回答者”进行某确定方案的随机抽样,对获取的子样本作“重点”访问. 这个方法相当于总体分为有回答与无回答两部分,如前一样假设这两部分所占的比例分别为 w_1 与 w_2 . 在调查方案中确定所取样本量为 n_0 , 每一次访问的费用为 c_0 , 而第一次访问后从第一部分中得到的回答为 n_1 , 对每一个这样的数据处理所需费用为 c_1 , $n_2 = n_0 - n_1$ 为无回答数. 例如用邮寄调查表的形式,收到回信的为“有回答”,在无回信的 n_2 个单元中抽选一部分用上门访问的方式进行第二次调查,通过努力最后又得到 $n'_2 = n_2/k$ 个数据. 设第二次数据的获得平均所需费用为 c_2 , 那末实际取得数据所需费用为

$$c_0 n_0 + c_1 n_1 + c_2 \frac{n_2}{k}. \quad (10.15)$$

这里 n_0 为预先确定, n_1 与 n_2 是随机的, $\frac{1}{k}$ 为第二次重点访问所占已知无回答者的比例数,是待定的某数. 因此,平均来说,所需费用为(10.15)式的期望:

$$O = c_0 n_0 + c_1 w_1 n_0 + \frac{c_2 w_2 n_0}{k}, \quad (10.16)$$

倘若待估计参数为 \bar{Y} , 可以用 \bar{y}_1 记第一次访问后得到的样本均值, 用 \bar{y}_2 记第二次重点访问所得数据的平均, 假如第二次访问的选取是随机的, 我们可以得到关于 \bar{Y} 的一个无偏估计:

$$\hat{\bar{Y}} = w_1 \bar{y}_1 + w_2 \bar{y}_2 = \frac{(n_1 \bar{y}_1 + n_2 \bar{y}_2)}{n_0}, \quad (10.17)$$

该估计的方差不难计算, 假设以 \bar{y}_2 作为 n_2 个无回答者有关数据的真正平均(当然这是无法获知的, 但在理论上它总归存在), 那末

$$\begin{aligned} V(\hat{\bar{Y}}) &= V[w_1 \bar{y}_1 + w_2 \bar{y}_2' + w_2(\bar{y}_2 - \bar{y}_2')] \\ &= V(w_1 \bar{y}_1 + w_2 \bar{y}_2') + V[w_2(\bar{y}_2 - \bar{y}_2')] \\ &\quad + 2\text{Cov}(w_1 \bar{y}_1, w_2(\bar{y}_2 - \bar{y}_2')) \\ &\quad + 2\text{Cov}(w_2 \bar{y}_2', w_2(\bar{y}_2 - \bar{y}_2')) \\ &= \left(\frac{1}{n_0} - \frac{1}{N}\right) S^2 + \frac{(k-1)w_2}{n_0} S_2^2, \end{aligned} \quad (10.18)$$

其中 S^2 为总体方差, S_2^2 为无回答部分的方差. 结合(10.16)及(10.18)式, 我们有可能按照实际情况来确定初始样本量 n_0 及无回答者中的再抽样比 k , 根据使 $O(V + s^2, N)$ 乘积达到最小的原则, k 的最佳选择为

$$k_0 = \sqrt{\frac{c_2(S^2 - w_2 S_2^2)}{S_2^2(c_0 + c_1 w_1)}}. \quad (10.19)$$

再根据 V 或者 O 确定的情况下, 分别从(10.18)与(10.16)式解出所需要的最初样本量 n_0 , 从(10.19)可知, 要知道 k_0 (从而解出 n_0), 必须还应知道 w_1, w_2 与 S_2^2 , w_1 与 w_2 常常可以根据资料或历史经验予以估计, 而 S_2^2 是无回答者部分的方差, 由于“无回答”, 当然较难知道它的大概. 显然 S_2^2 不能用 S^2 来代替, 因为无回答部分常常有其自己的特性, 这的确增加了估计 k_0 与确定 n_0 的难度, 但不管怎样, 这种方法本身确实对于“无回答”所引起的误差产生校正的效果.

(二)常常关心的是, 如果只进行一次调查, 由于无回答而产生的误差如何校正. Polize-Simmons(1949, 1950)对总体均值的建议有一定的启发性.

假定所有访问均是在星期日以外六个晚上进行. 对于每一个在访问中遇到的被调查者附加询问其在前面五个晚上(不包括星期日)是否在家. 根据他所回答的在家天数 t 就不难获得他在家的频率 ω 的估计: $\hat{\omega} = (t+1)/6$.

调查的结果可以因 t 的取值不同 ($t=0, 1, 2, \dots, 5$) 而划分为六个部分, 每个部分含有 n_t 个调查数据, t 越大, π 就越大, 该组入样的可能性就越大, 因此该部分的均值 \bar{y}_t 在估计总体均值的过程中将赋予与 π 相对应的权, 这与不等概率抽样时的均值估计有类同的意思, 于是我们将样本均值 \bar{y} 调整为 Polize-Simmons 估计:

$$y_{ps} = \frac{\sum_{t=0}^5 6n_t \bar{y}_t / (t+1)}{\sum_{t=0}^5 6n_t / (t+1)} = \frac{\sum_{t=0}^5 n_t \bar{y}_t / (t+1)}{\sum_{t=0}^5 n_t / (t+1)}. \quad (10.20)$$

这种校正直观上有个合理的假设: 某些感兴趣的指标 (例如生活费用的平均年支出) 与被调查者是否容易找到这个因素有较大的相关性, 如果笼统地采用所得数据的均值, 很可能偏于突出哪些容易找到的对象的相应数据的影响, 而掩盖了另一部分人相应数据的作用. 采用 Polize-Simmons 的加权平均在一定程度上校正了这一偏差. 不容置疑的一点是: 由于采用加权均值代替了一般均值, 且这种权数也是通过估计得来的, 因此我们将不得不付出增大估计量方差的代价. 但节省时间、费用是这种校正的优点, 因为它不需要作再访问.

§ 10.2 调查误差

现在我们回到处理数据的问题, 而不去考虑无回答的影响. 假如通过抽样调查得到了 n 个数据, 由于计量工具的不够精确, 以及调查员的工作中的某些失误等等, 都可能造成调查数据存在一定的误差. 甚至还有可能某些数据存在着“伪劣”现象, 它非但不能提供有关的信息, 反而干扰了在处理数据时作出的推断的正确性. 这一类所谓“伪劣”数据常常因为被调查者不了解情况, 或者甚至所提出的问题涉及到一些敏感问题, 或是由于某些被调查对象出于某种动机而提供虚假材料. 在本节中我们将简略地讨论这两种出现误差的情况.

10.2.1 调查误差的数学模型

在理论上我们可以对第 i 个单位进行 k 次重复调查并作计量, 令 $y_{i\alpha}$ 为第 α 次重复计量中所得数据, 它可以表示为

$$y_{i\alpha} = \mu_i + e_{i\alpha}, \quad (10.21)$$

其中 μ_i 为第 i 单位某指标的真实数据, 而 $e_{i\alpha}$ 则为对第 i 单位第 α 次计

量时的观测误差,一般讲来, $e_{i\alpha}$ 将遵从一个分布,譬如一般的测量误差依照误差理论服从一个正态分布. 设 $Ee_{i\alpha} = \beta_i$, 假如 $\beta_i = 0$, 那末只要对第 i 个单位多重重复计量几次, 根据大样本理论所得的平均值就可在相当程度上接近于真值 μ_i . 如果 $\beta_i \neq 0$, 那末就表示在观测中出现了一定的系统误差. 需要指出: β_i 以及误差 $e_{i\alpha}$ 的方差 σ_i^2 不仅与计量工具有关, 还常常与所调查的指标有关. 尤其是在以人为总体的某些指标的调查中, 出于政治、经济等诸方面的因素, β_i 常常不为零.

对于固定的 i , 显然偏差 β_i 是个常量, 但是随着 i 的不同, β_i 也不同. 若 $E\beta_i = \beta$, 则 β 称之为所有调查单位的常数偏差. 变量 $(\beta_i - \beta)$ 也将遵从一个分布. 偏差的这一组成与真值 μ_i 有关, 例如在实际观测中高估或低估了真值 μ_i . 若记 $d_{i\alpha} = e_{i\alpha} - \beta_i$, 则对于每一固定的 i , (10.21) 式可以表示为

$$y_{i\alpha} = \mu_i + \beta + (\beta_i - \beta) + d_{i\alpha}. \quad (10.22)$$

按照 $d_{i\alpha}$ 的定义, 它表示了对每一固定的 i , 进行观测时误差中的波动部分, 它与 $e_{i\alpha}$ 具有同样形状的分布, 只不过其期望为 0.

最简单的情况是: 对于所有的 i, α , 观测值 $y_{i\alpha}$ 是独立获取的. 但在社会经济的抽样调查中未必达到这一点. 同一单位的不同次观测以及不同单位之间的观测变量都有可能在一定程度上相关. 这往往由人的因素、环境的变化等等原因所造成, 尤其是抽样调查是对一些社会经济现象的观测, 某些事物之间本身就存在着的相关性是不以调查者的意志所转移的. 例如观察证券市场的股价变化, 每一种股票的若干次价格记录是有一定的相关程度, 而各种股票的股价之间也明显地存在着相关性, 某些股票的股价上升会引起其他有些股价的波动. 另一种稍稍不同的模型描述是以 $d_{i\alpha}$ 及

$$\mu'_i = E(y_{i\alpha}, i) = \mu_i + \beta_i \quad (10.23)$$

表示的. μ'_i 实质上是对第 i 个单位若干次重复计量的平均. 因此模型可写成:

$$y_{i\alpha} = \mu'_i + (e_{i\alpha} - \beta_i) = \mu'_i + d_{i\alpha}. \quad (10.24)$$

$d_{i\alpha}$ 是第 i 个单位(或被调查者)对调查作的若干次回答所产生的差异, 称之为回答离差(response deviation), 若记 $\bar{\mu}' = \frac{1}{n} \sum_{i=1}^n \mu'_i$ 而变量 μ'_i 在总体中的均值设为 μ' , μ 为总体均值的正确值, 那末

$$y_{i\alpha} - \mu = d_{i\alpha} + (\mu'_i - \mu') + (\mu' - \mu). \quad (10.25)$$

(10.25)式关于样本取平均:

$$\bar{y}_a - \mu = \bar{d}_a + (\bar{\mu}' - \mu') + (\mu' - \mu). \quad (10.26)$$

于是得到均方误差公式:

$$\text{MSE } y_a) = V(\bar{d}_a) + V(\bar{\mu}') + (\mu' - \mu)^2 + 2\text{Cov}(\bar{d}_a, \bar{\mu}') \quad (10.27)$$

(10.27)式右端各项分别称作回答方差、抽样方差与偏倚平方。而第4项的协方差,由于在模型中 $E(d_{ia}, i) = 0$, 因此该项一般取值为零。

以下研究调查误差数字模型中各组成部分的影响以及误差方差的评估:

(一)所有单位的常数偏差 β , 如果存在的话, 那末对样本均值等显然有一定影响。而对于相应方差, 由于其形式为 $(y_i - \bar{y})$ 的平方和, 在每一次中常数偏差恰被抵消, 因此方差估计不受常数偏差的影响。对于其他常见的一些估计量, 上述结论几乎也成立。即常数偏差 β 的存在使估计量也产生偏差, 但估计量的误差方差并无影响。这一点是不难理解的, 因为我们所处理的估计量常常可以近似地表示成观测值的某种函数的均值形式。

但是, 仅从样本资料本身不可能查出常数偏差, 因为每一个样本几乎都有一个定量的移动, 只能使人们对总体的有关指标产生误导而根本无法觉察。在这种情况下, 历史资料、经验常识等也许可以帮助我们区分出较显著的常数偏差 β 。

(二)考虑最简单的情况, 即所取样本中调查误差是互不相关的, 由 (10.27)式, 易知

$$V(\bar{y}_a) = V(\bar{d}_a) + V(\bar{\mu}'). \quad (10.28)$$

具体实施时, 先将某样本经若干次重复调查所得的数值平均, 然后再对不同简单随机样本取平均。令 $V(d_{ia}) = \sigma_i^2$, 对于具体一组样本 (y_1, \dots, y_n) 有

$$E(\bar{d}_a^2 | (y_1, \dots, y_n)) = \frac{1}{n^2} \sum_i^n \sigma_i^2. \quad (10.29)$$

考虑到每个单位入样的可能性为 $\frac{n}{N}$, 则有

$$V(\bar{y}_a) = \frac{1}{nN} \sum_i^N \sigma_i^2 + \left(\frac{1}{n} - \frac{1}{N} \right) \frac{\sum_i^N (\mu_i' - \mu')^2}{N-1} \quad (10.30)$$

$$\triangleq \frac{1}{n} \sigma_d^2 + \frac{1-f}{n} S_{\mu'}^2. \quad (10.31)$$

如果采用通常简单随机抽样估计方差的方法:

$$v(y_a) = \frac{1-f}{n} S^2 = \frac{1-f}{n} \cdot \frac{\sum_i (y_{ia} - \bar{y}_a)^2}{n-1}, \quad (10.32)$$

由于

$$y_{ia} - \bar{y}_a = (d_{ia} - d_a) + (\mu'_i - \mu'), \quad (10.33)$$

于是

$$Ev(\bar{y}_a) = \frac{1-f}{n} \sigma_d^2 + \frac{1-f}{n} S_{\mu'}^2, \quad (10.34)$$

假如 $f = n/N$ 相当小, 与(10.31)式相比, 我们可以认为 $v(\bar{y}_a)$ 几乎是 $V(\bar{y}_a)$ 的无偏估计.

(三) 考虑了 d_{ia} 为互不相关的情况之后, 我们当然要研究样本中各单位 d_{ia} 之间相关的情况, 此时

$$\bar{d}_a^2 = \frac{1}{n} \left(\sum_i d_{ia}^2 + \sum_{i \neq j} d_{ia} d_{ja} \right), \quad (10.35)$$

因此

$$\begin{aligned} V(d_a) &= E(\bar{d}_a^2) \\ &= \frac{1}{n} \sigma_d^2 + \frac{n-1}{n} E(d_{ia} d_{ja}) \quad (i \neq j). \end{aligned} \quad (10.36)$$

定义样本内相关系数为

$$\rho_w = E(d_{ia} d_{ja}) / \sigma_d^2, \quad (10.37)$$

代入(10.36)式, 得

$$V(\bar{d}_a) = \frac{\sigma_d^2}{n} [1 + (n-1)\rho_w]. \quad (10.38)$$

式中 $V(\bar{d}_a)$ 称为总回答方差, σ_d^2/n 称为简单回答方差, 而 $(n-1)\rho_w\sigma_d^2/n$ 称为总回答方差中的相关分量.

假如 $\text{Cov}(d_a, \bar{\mu}') = \gamma$, 则(10.31)式成为

$$V(\bar{y}_a) = \frac{\sigma_d^2}{n} [1 + (n-1)\rho_w] + \frac{1-f}{n} S_{\mu'}^2. \quad (10.39)$$

采用通常的估计方差形式(10.32), 则其期望为

$$Ev(\bar{y}_a) = \frac{1-f}{n} [\sigma_d^2(1-\rho_w) + S_{\mu'}^2]. \quad (10.40)$$

对于许多种调查误差, ρ_w 很可能是正的, 即当某单位的调查误差呈现正差异的话, 常常引起另一单位也具有正差异的调查误差. 在这种情况下, 利用 $v(\bar{y}_a)$ 来估计 $V(\bar{y}_a)$ 常常偏低.

样本之间调查误差的相关问题常见于“调查员问题”，尤其是调查员处理的是一些涉及意见性或判断性方面的定性内容时，更容易产生样本之间调查误差的相关。

对上述三种情况作一归纳，得到

$$\text{MSE}(\bar{y}_a) = \frac{1}{n} \{S_{\mu'}^2 + \sigma_a^2[1(n-1)\rho_w]\} + \beta^2, \quad (10.41)$$

其中 $\mu'_i = \mu_i + \beta_i$ 。

在公式(10.41)中，随着样本量 n 的不断增大， $S_{\mu'}^2/n$ 及 $\sigma_a^2(1-\rho_w)/n$ 随之越来越小。但是另外两项 β^2 与 $\sigma_a^2\rho_w$ 并不随 n 的变化而变化。在实际操作中，并非如此简单。假如样本容量 n 相当大，我们在大规模的调查中很可能在具体计量的方式中有所改变，因为在费用、时间消耗以及数据处理方面， n 的大小不同对调查很有影响。这种计量方式的改变自然影响了 β 与 ρ_w 的数值。凭直观想象， β 与 ρ_w 的这种变化一般比起 n 来是较缓慢的。因此，在大的样本量中，这两项成了 MSE 的主要组成部分，相对地，抽样方差反而显得不太重要。此时用 MSE 以评估估计的精确性就有些欠妥。

10.2.2 几种处理方法

由于调查误差的产生引起估计量的偏倚以及影响对估计量的真正正确性作出判断。因此对调查误差的研究引起了人们的注意。其中可能出现的问题是各种不同的调查所产生的调查误差是不一样的，要处理好“调查误差”的影响需因地制宜。

最理想的方法是完全取得正确数据 μ_i ，但是这种在理论上行得通的事情，在实际中往往未必遂人心愿。因为它涉及到费用、时间等问题，而且在计量过程中，很少有哪些器具使之不产生任何误差。

在无法保证能取得正确数据 μ_i 的情况下，我们只得另辟途径，或者用更正确可靠的方法重新计量，或者利用横向或纵向的比较，（即比较两个总体的同一指标，或者比较同一总体不同时期的同一指标等等），从而对调查计量偏差 β 至少有个粗略的估计。再接下去的处理就是对样本估计量的方差的各组成部分（例如抽样方差及回答方差等）作出数量上的估计。

一、随机子抽样方法

假如有 K 个调查员对某总体进行一次抽样调查，规定每人完成 m 个

单位的计量。为了评估这次调查的质量,通常所采用的方法是从这 K 个调查员中随机抽取 k 个,再组织 k 个具有同样训练素质的调查员对他们各自完成的调查单位重新调查。

现在考虑某一对调查员所调查的数据,设由他们调查第 i 个单位后所得的数据分别记为 $y_{i1}, y_{i2} (i=1, 2, \dots, m)$, 按数学模型:

$$y_{it} = d_{it} + \mu'_i \quad (t=1, 2), \quad (10.42)$$

正如在第9章中所介绍的那样, y_{i1} 与 y_{i2} 之间的差的平方提供了该单位调查误差方差的信息, 将这对调查员所调查的单位得到的数据差平方加以平均:

$$E\left\{\frac{\sum_{i=1}^m (y_{i1} - y_{i2})^2}{2m}\right\} = \frac{\sigma_{d1}^2 + \sigma_{d2}^2}{2} - \text{Cov}(d_{i1}, d_{i2}). \quad (10.43)$$

现在提出如下假设:

(1) 关于同一单位的回答误差 d_{i1} 与 d_{i2} 不相关;

(2) 第1次调查人员的简单回答方差 σ_{d1}^2 与再调查人员的简单回答方差 σ_{d2}^2 相等。

上述假设(1)、(2)在通常情况下具有一定的合理性。因为我们总是假定前后两次调查人员的调查是独立进行的,这一点保证了(1)的成立。而两位调查人员具有同样的训练素质则保证了假设(2)的成立。

在假设(1)、(2)成立情况下,公式(10.42)提供了 σ_{d1}^2 的一个良好估计,由于 $\sum_{i=1}^m (y_{i1} - y_{i2})^2 / 2m$ 是仅对一对调查员而言,只要将 k 对调查员相应的公式相加再平均就成为 σ_{d1}^2 的估计量。

当然,也存在着假设不成立的情况,例如被调查者在第二次调查中仅仅依靠回忆第一次回答的内容,而不是“重新独立”地考虑回答的内容,此时显然获取了正的协方差 $\text{Cov}(d_{i1}, d_{i2})$, 这样利用 k 个 $\sum_{i=1}^m (y_{i1} - y_{i2})^2 / 2m$ 的平均去估计 σ_{d1}^2 会发生“低估”现象。

为了利用随机子抽样方法对调查质量作出恰当评估,尽量使假设(1)、(2)成立是值得的,就组织者而言,不让第二个调查员了解第一次调查的结果也许是有益的。

二、交叉子样本方法

除了简单回答方差之外,我们还需要对总回答方差中的相关分量有所了解。由数理统计学中方差分析的知识,为了分解出方差的各种成分,

最好是将方差估计公式中的平方和进行类似于组内离差与组间离差等部分。在抽样调查中,相应的较好方法无非是将样本随机分为若干组,然后由不同的调查员独立地对每组进行调查,这就是所谓的“交叉随机子抽样方法”,具体实施如下:

n 个待调查的样本单位随机地分为 k 个子样本, 每个含 $m = n/k$ (假如 n 可以被 k 整除的话) 个单位, 假定这 k 个子样本的单位之间不存在调查误差的相关性, 这一点在许多场合是容易做到的。不然的话, 在划分 k 组时应将这个因素考虑进去。指派 k 个调查员分别对这 k 个子样本进行调查, 调查一般是独立执行的。因此, 不同调查员之间不存在调查误差的相关这一假设是合乎情理的。现在建立数学模型如下:

$$y_{ija} = \mu_{ij} + d_{ija}, \quad (10.44)$$

其中 i 表示第 i 个子样本 (或第 i 个调查员), j 表示该子样本中第 j 个单位。在第 i 组内, 由 (10.38) 式得

$$V(y_{ia}) = \frac{1}{m} \{S_{\mu}^2 + \sigma_d^2 [1 + (m-1)\rho_w]\}. \quad (10.45)$$

这里的 ρ_w 是指同一调查员所得 d_{ija} 之间的相关。

由于各不同子样本中调查误差的独立性, 易得

$$\begin{aligned} V(\bar{y}_a) &= \frac{1}{k} V(y_{ia}) \\ &= \frac{1}{n} \{S_{\mu}^2 + \sigma_d^2 [1 + (m-1)\rho_w]\}. \end{aligned} \quad (10.46)$$

如 (10.32) 式所述, 对 $V(\bar{y}_a)$ 的估计常采用 $\sum (y_{ia} - \bar{y}_a)^2$ 乘上某一常数因子的形式, 在交叉随机子抽样模型中, $\sum (y_{ia} - \bar{y}_a)^2$ 变成 $\sum_i \sum_j (y_{ija} - \bar{y}_a)^2$, 则有

$$\begin{aligned} S^2 &= \sum_{i=1}^k \sum_{j=1}^m (y_{ija} - \bar{y}_a)^2 \\ &= \sum_{i=1}^k \sum_{j=1}^m (y_{ija} - \bar{y}_{ia})^2 + m \sum_{i=1}^k (\bar{y}_{ia} - \bar{y}_a)^2 \\ &\triangleq S_w^2 + S_b^2. \end{aligned} \quad (10.47)$$

显然, S_w^2 的自由度为 $k(m-1)$, S_b^2 的自由度为 $k-1$ 。经简单的期望运算, 得

$$ES_w^2/k(m-1) = S_{\mu}^2 + \sigma_d^2(1-\rho_w), \quad (10.48)$$

$$ES_b^2/(k-1) = S_{\mu}^2 + \sigma_d^2[1 + (m-1)\rho_w]. \quad (10.49)$$

因此, 在本模型中可以利用交叉随机子样本平方和 $S_b^2/km(k-1)$ 作为 $V(\bar{y}_a)$ 的无偏估计。而从 (10.48)、(10.49) 式可以看到

$$E\left[\frac{S_{b1}^2/(k-1) + S_w^2/k(m-1)}{m}\right] = \sigma_{\theta\omega}^2. \quad (10.50)$$

这蕴含了总回答方差的相关分量可以利用交叉随机子抽样估计量 $[S_{b1}^2/(k-1) + S_w^2/k(m-1)]/m$ 。当然也可以估计出相关分量在总回答方差中所占的份量。

10.2.3 数值异常情况

在抽样调查中,数据方面的缺陷除了无回答、调查误差等以外,还存在着数值异常现象。所谓“数值异常”是指调查所获得的数据超出正常范围之外。这种现象是由一些重大事件或某些异常因素引起的。数值异常现象大致分为两种:一种是数据虽属异常,但却是真实的。例如在调查部分地区经济现状或该年度生产总值时,恰逢某入选样本在调查期发生了严重自然灾害,此时我们手中获得的资料显然呈“异常”,但它在事实上反映了问题的本来面目。这一类异常的数据往往在第一次调查时就会被发觉,或者通过再调查核实时被发觉,并找到成因。第二种异常数据是人为地制造的,例如偏离实际的虚报,应付任务式的编造等等。这种数据属于“伪劣”数据,会对抽样推断的结果产生很大的偏差及影响。

关于如何发现与判断数据是否异常的问题,通常只有比较样本数据的整体变化才有可能确定,有时也利用历史样本以及经验进行判断。例如调查小麦的亩产量,发现个别数据值为5000公斤(假设),从纵向(历史上小麦亩产量)及横向(附近地区所得小麦亩产量的抽样数据)作比较,它均表现为“突出值”,我们就有理由怀疑该数据的真实性,并作出删去或再调查核实的决策。在实际的抽样调查中,由于我们调查的数据有些涉及到该单位的机密等事宜,有可能发生人为的虚报假报。故我们在一次调查中常常不是只调查一个指标而是要调查若干个指标,注意到在社会经济中许多指标之间存在着一定的相关。从以前的或其他的调查中,我们常常可获知这些指标的相关程度的数量上的估计(至少是粗略的估计)。这种知识对我们及时发现一些人为的故障可提供帮助。我们曾在1993年协助有关机构对企事业的工资、人员、福利等作过抽样调查,发现某些单位所填报表的各项指标中有明显不符合常识的相互关系,因此断定这些单位所提供的数据属“虚伪”的,如果不作处理,是无法采用的。这个简单的事实启示了我们,如果巧妙地设计调查表,尤其是巧妙地插入一些相关的指标,有时能及时发觉某些人为的虚假编造。当然要真正地杜绝这

种人为的虚假数据,应当借助于“统计法”的立法及健全.

对于异常数据的处理,无非是删除或者在可能的情况下作再调查.对“伪劣”数据必须删除!因为它对抽样推断起着破坏性的影响,但对于第一类的异常数据采用删除方式要慎重,因为它毕竟反映了一定的信息.对于数值异常问题的研究正在引起国内外有关学者的重视.

§ 10.3 敏感性问题的调查

10.3.1 敏感性问题的调查与随机化回答

在社会经济调查中,有时提出的一些问题是属于敏感性的或高度私人绝密的内容.例如在调查科技人员的流向及有关意愿时,被调查者出于种种原因,不愿在流动之前坦露自己的意向,以免在原工作单位造成不必要的麻烦.如果我们的调查内容仅仅限于是否想离开原单位,而且我们能够设计一种方案,做到被调查者可以作出真实回答又能保守私人秘密,那末这个问题就得到圆满解决.

Warner(1965)曾针对仅有“是”或“否”两种回答的调查(目的是获得总体中“是”的比例)设计了一种随机化装置达到了上述目的.基本思路如下:

对于 n 个被调查者中的每一个以概率 P 及 $(1-P)$ 提出两个截然相反的问题,例如“我赞成某事”或“我不赞成某事”,Warner 装置的巧妙之处在于调查人员无法知道被调查人员在回答哪个问题,要做到这一点并不难,例如只要准备几张折叠白纸(折叠以后外形完全一致),以 P 比 $(1-P)$ 的相对比例在每张纸内写上提出的两个问题之一.被调查者随机地摸取一张纸回答,但调查人员无权查看纸条上的问题.对所提出的问题,被调查人员只有两种选择“是”或“否”,他可以将红球(表示“是”)或白球(表示“否”)投进一个密封的口袋中,整个投球过程也是调查人员所看不到的.如果向被调查者讲清 Warner 方案的具体作法以及严格地依照此方案进行调查的话,那末就容易使被调查者确信他或她参加了这次调查但绝不会泄露自己在这个敏感性问题中的态度.

假如所得红球为 m 个,那末总体中回答“是”的比例 ϕ 显然可以用 $\hat{\phi} = m/n$ 作为估计量.而概率统计常识告诉我们,总体中“赞成某事”的比例 π 与 ϕ 及 P 具有下述简单的关系式:

$$\phi = P\pi + (1-P)(1-\pi) = (2P-1)\pi + (1-P), \quad (10.51)$$

式中 P 是调查者自己在方案制定中所确定的已知数, 因此, 按照关系式 (10.51) 可以得到估计量:

$$\hat{\pi} = \frac{[\hat{\phi} - (1-P)]}{2P-1} \quad (P \neq 1/2). \quad (10.52)$$

注意到 $\hat{\phi}$ 实质上是二项分布中成功概率 P 的极大似然估计与无偏估计, 因此 $\hat{\pi}$ 当然是 π 的极大似然估计与无偏估计, 其方差不难求得为:

$$V(\hat{\pi}) = \frac{\phi(1-\phi)}{n(2P-1)^2}. \quad (10.53)$$

如果将 $(1-\phi)$ 写成

$$1-\phi = (2P-1)(1-\pi) + (1-P), \quad (10.54)$$

则可求得

$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \quad (10.55)$$

$V(\hat{\pi})$ 中的第一项表示了假定对所有 n 个被调查者都直接提问“是否赞成某事”, 并且这 n 个被调查者都如实地回答了这个敏感性问题后, 从而得到 π 的估计 $\hat{\pi}$ 所应具有 的方差. 而 (10.55) 式右端的第二项一般说来 (除去 π 非常接近于 $1/2$, 而 $P > 0.85$ 又同时成立的情况) 比第一项大得多. 这表明了使用 Warner 方案后得到 π 的估计 $\hat{\pi}$ 一般来说精确度很差, 这一点不难想象, 因为调查员看到的红球数由于提问的随机性而不能知道这是不是“赞成某事”的人数或者“不赞成某事”的人数. 但由于所提问题的敏感性, 这个粗糙的估计总比被调查者拒绝回答或给予一个“伪劣”性质的回答要强得多. Warner 在理论上还证明了在涉及敏感性问题的调查中, 他的方法比直接提问调查的均方误差 (MSE) 小.

10.3.2 Simmons 问题

Warner 方法的成功与否取决于被调查者确信自己的态度没有公开暴露从而愿意采取合作的态度. Simmons 提出 (Horvitz, Shah and Simmons, 1967) 如果将第二个问题改为与第一个问题毫无关系, 也许可以改进被调查者的合作程度. 例如第一个陈述仍为“我赞成某事”, 若将第二个陈述改为“我的生肖属狗”, 调查过程仍采用 Warner 的随机化问答. 明显地, 比起前一节来, 这里第二个问题几乎与第一个问题无关, 另一个不同点是调查者对总体生肖属狗的比例可能是清楚的, 比如大概为 $\frac{1}{12}$, 即对第二个陈述回答“是”的人在总体中的比例 π^* 为已知, 此时

$$\varphi = P\pi + (1-P)\pi^*, \quad (10.56)$$

因此, 感兴趣的比率 π 的极大似然估计量为:

$$\hat{\pi} = \frac{[\hat{\varphi} - (1-P)\pi^*]}{P}, \quad (10.57)$$

相应方差为

$$V(\hat{\pi}) = \frac{\varphi(1-\varphi)}{nP^2}. \quad (10.58)$$

有趣的一点是: Dowling 与 Shachtman (1975) 证明了当 $P > \frac{1}{3}$ (注意 $P \neq \frac{1}{2}$) 时, 不管 π 与 π^* 如何, 则用 Simmons 方法得到的 $\hat{\pi}$ 的方差将小于用 Warner 方法估计量 $\hat{\pi}$ 的相应方差, 也就是说, 同样是极大似然估计, Simmons 方法在 $P > \frac{1}{3}$ ($P \neq \frac{1}{2}$) 情况下要精确一些.

倘若对第二个问题(或陈述)回答“是”的人在总体中所占有的比例 π^* 无法知道, 那末我们将面临两个未知的 π 及 π^* , 最好的办法当然是通过两组样本(容量分别为 n_1, n_2)来解决, 假如这两组样本中提出敏感性问题比例分别为 P_1 与 P_2 , 而 φ_1 与 φ_2 分别表示相应于 P_1 与 P_2 的在总体中回答“是”所占的比例, 于是有

$$\varphi_1 = P_1\pi + (1-P_1)\pi^*, \quad (10.59)$$

$$\varphi_2 = P_2\pi + (1-P_2)\pi^*, \quad (10.60)$$

这样可以得到

$$\hat{\pi} = \frac{\hat{\varphi}_1 - (1-P_2) - \hat{\varphi}_2(1-P_1)}{P_1 - P_2}, \quad (10.61)$$

其方差为 d ,

$$V(\hat{\pi}) = \frac{1}{(P_1 - P_2)^2} \left[\frac{\varphi_1(1-\varphi_1)(1-P_2)^2}{n_1} + \frac{\varphi_2(1-\varphi_2)(1-P_1)^2}{n_2} \right]. \quad (10.62)$$

其实, 从(10.59)、(10.60)式同样也能解出 π^* 的估计量. 这意味着通过两组样本调查, 我们可以同时对两个无关的敏感性问题作出估计(如果需要的话). 但是在实践中事情并非那么简单, 因为要求被调查者同时对两个或更多的敏感性问题表示自己的态度, 容易引起怀疑从而出现拒绝回答或虚报回答等不合作现象.

10.3.3 数值例子

为了解目前一些青年学生对某些课程感兴趣的程度, 可以通过抽样

调查的手段解决。但是必须注意到我们所提出的问题常常是个敏感性问题,因为它涉及到有些学生对该课程的任课老师或其他一些问题的忌讳,尤其是对有些政治课程的提问,更带有敏感性质。

例如我们的陈述为“我对社会主义思想教育课感兴趣”以及“我对社会主义思想教育课不感兴趣”,对此敏感性的问题我们采用 Warner 方案处理,预先指定的 $P = \frac{4}{5}$,在接受调查并作出明确回答的 320 人中(由于方案事先解释得清楚且执行得认真,我们认为这 320 个人都真实地回答了问题),统计最后结果,回答“是”的人为 156 人,按照计算公式得 $\hat{p} = 0.479166 \approx 0.48$, 及 $\sqrt{V(\hat{p})}$ 约为 0.047。

本例可以采用 Simmons 形式处理,但应注意到第二个陈述的合理选择,所谓“合理选择”,需要注意下述两点:

(一)该陈述不宜采用“暴露”性问题,例如“我是男性”这样的陈述虽然与第一陈述“我对社会主义思想教育课感兴趣”几乎没有太大内在联系,但在抽样过程中性别问题本身已经暴露,如果采用这样类型的陈述很可能使被调查者不愿确信调查的保密性,从而引起调查结果未必全部真实,这是我们所不愿看到的事情。

(二)第二陈述应尽量与第一陈述无关。这是 Simmons 问题本身的一个关键。这方面万一发生差错将使数学模型不适宜具体调查,那么就不可能利用(10.56)、(10.57)式对待估的比例及其方差作出恰当的估计。比如我们想了解青年中参与赌博行为的人所占的比例,显然,这也是个敏感性问题,假定我们能使被调查者确信调查方法的保密是绝对可靠的,那么我们提出的第一个问题毫无疑问为“我有赌博行为”,鉴于该问题的敏感性相当突出,似乎采用 Simmons 形式为佳,因为这样可以使调查者更为放心。倘若我们的第二陈述采用“我参加娱乐活动是为了劳逸结合”,这两个陈述似乎有一定关联,因为有些青年的确认为赌博是“玩玩而已”的娱乐活动,这就有可能干扰我们的推断。我们不妨采用一个与第一陈述毫无关系的陈述,比如“我的生肖属狗”等。

10.3.4 具多项选择的敏感性问题的调查

前面所讨论的有关敏感性问题调查建立在回答只有“是”与“否”两种选择的基础上。在实际抽样调查中,有些敏感性问题的回答可以有若干种选择。设想某地区将从甲、乙、丙三人中推选出一名代表,该地区的每

个成人在投票时有 4 种互不相容的选择, 甲、乙、丙或弃权. 倘若在事先抽样摸底中, 一般人们不愿过早表明自己的态度, 那么该问题就是一个有 4 项选择的敏感性问题. 本段主要叙述具多项选择的敏感性问题的调查方案设计以及有关数学模型.

设 S 是一个敏感性问题, 对于 S 的回答有 k 种选择, 记为 A_1, A_2, \dots, A_k . 假定大小为 N 的总体中对 S 取回答为 A_i 的人数为 $N_i (i=1, 2, \dots, k)$, 我们关心 $P_i = N_i/N (i=1, 2, \dots, k)$.

方案设计仍采用抓阄法, 在 m 张大小、质地同样的折叠白纸条上分别标上号码 $0, 1, 2, \dots, k$, 其个数分别为 $m_0, m_1, m_2, \dots, m_k, \sum_{i=0}^k m_i = m$. 现在设想被调查者随机地摸取一张纸条, 其中所标的号码仅有被调查者知道, 如果他摸到的标号为 0, 那末他必须依照自己真实的想法回答 $A_i (i=1, 2, \dots, k)$, 具体作法是将手中唯一红球投到 k 个匣子中标有 A_i 的那一个, 如果他摸到纸条标号为 $i (i=1, 2, \dots, k)$, 那末他必须将手中唯一的红球投入标有 A_i 的匣中而不管他自己的真实态度如何. 整个投球过程均在调查者无法观看的情况下进行. 设 f_0 为摸到标号为 0 的纸条的概率, f_i 为肯定不是 0 标号情况下, 标号为 $i (i=1, 2, \dots, k)$ 的概率, 显然它是个条件概率, 具体计算如下:

$$f_0 = m_0/m, \quad (10.63)$$

$$f_i = m_i/(m - m_0). \quad (10.64)$$

以 n 表示总样本人数, 而 n_i 则表示在标有 A_i 的匣子中的球数, $\sum_{i=1}^k n_i = n$ (n_1, n_2, \dots, n_k) 是一组随机变量, 假如以 q_i 表示在本方案设计之下总体中每一个人选择 A_i 的概率. 那末显见向量 (n_1, n_2, \dots, n_k) 服从参数为 q_1, q_2, \dots, q_k 的多项分布:

$$P\{X_1 = n_1, X_2 = n_2, \dots, X_k = n_k\} = \frac{n!}{n_1! n_2! \dots n_k!} q_1^{n_1} q_2^{n_2} \dots q_k^{n_k}, \quad (10.65)$$

其中 X_i 表示随机化回答中取 A_i 的随机人数, 而

$$q_i = f_0 P_i + (1 - f_0) f_i, \quad (i=1, 2, \dots, k), \quad (10.66)$$

从 (10.66) 式解得

$$P_i = \frac{q_i - (1 - f_0) f_i}{f_0} \quad (i=1, 2, \dots, k). \quad (10.67)$$

显然 q_i 可以用 $\hat{q}_i = n_i/n$ 估计, 因此

$$\hat{P}_i = \frac{\hat{q}_i - (1-f_0)f_i}{f_0} \quad (i=1, 2, \dots, k). \quad (10.68)$$

利用多项分布(10.65)式的性质, X_i 服从成功概率为 q_i 的二项分布, 因此, \hat{P}_i 是 P_i 的极大似然估计及无偏估计, 其方差为

$$V(\hat{P}_i) = \frac{q_i(1-q_i)}{nf_0^2} \quad (i=1, 2, \dots, k). \quad (10.69)$$

当 $i \neq j$ 时, 我们还可以得到两个估计 \hat{P}_i 与 \hat{P}_j 的相关系数:

$$\rho(\hat{P}_i, \hat{P}_j) = -\sqrt{\frac{q_i q_j}{(1-q_i)(1-q_j)}} \quad (i \neq j). \quad (10.70)$$

第 11 章

案例分析

§ 11.1 引言

正如我们在第 1 章引论中所说的, 抽样调查历来是应用最为广泛的数理统计方法之一。随着我国改革开放的不断深化, 社会主义市场经济体制逐渐建立, 人们对各种信息的需求日益强烈, 抽样调查被看作是一种快速、经济而有效的获取资料的重要手段。为适应这种形势的变化, 我国已一改过去以统计报表制度为基础的全面调查(普查)这种单一的调查形式, 初步形成了抽样调查与全面调查互为补充、相辅相成的收集各种统计信息资料的格局。而且随着形势的发展, 抽样调查必将愈显重要而居主导地位。

最近 15 年来, 我国各级政府部门(包括国家统计局及其他主管部门)、经济实体及学术研究机构以至许多新闻单位与民间机构进行了数以千百计的目标多样、规模不等的抽样调查项目。根据抽样调查获得的数据与结论不断的见诸于各类报刊和其他传播媒介以及内部研究报告等。就应用领域看, 这些项目几乎包括了社会、经济、文化、教育、卫生和科学研究等各种领域。但在这些项目中, 并不是每项都是成功的。事实上, 一项抽样调查, 除了必要的经费和组织保证外, 它的成功与否主要取决于它的设计与分析, 而这一点正是许多人所容易忽略的。不懂抽样调查理论与方法的人当然不理解设计的重要性。即使学了一些抽样调查理论的人在遇到实际问题时也往往感到束手无策, 不能保证将一项实际调查设计好。这是因为, 诸如: 工作人员缺少实际经验, 对于实际中不同项目的调查, 由于其目的与对象不同, 抽样单元与抽样框的形式各异, 投入的经费及人力相差悬殊等等。因此, 对具体设计与分析的要求有很大差别, 更不要说它必然受到种种主客观条件的限制了。一项实际抽样调查不可能只采用一种简单的抽样与分析方法, 而往往是多种抽样方法有机的组合。对于一项大规模的, 例如全国性的调查尤其如此。

为使读者对运用抽样调查方法有感性的了解。在本章中, 我们选择

了十来项实际案例进行研究与分析。每个案例均介绍了调查的背景、目的及具体的抽样设计,大部分还包括了数据分析方法,重点在于总体目标量的估计与方差估计。有些案例还包括对结果的精度分析和其他分析方法等。对每个案例,我们都加了评注,进一步引导读者理解设计思想,并指出(如果存在的话)其中不足以及在条件许可的情况下可以改进的地方。半数左右的案例取自作者本人的实践,但也包括其他一些项目,特别是国家统计局制定的两项定期的有关农产量抽样调查和人口变动量抽样调查以及卫生部的国家卫生服务总调查,全国妇联的中国妇女社会地位调查。作者谨向有关部门及设计者表示衷心的感谢。

为保持每个案例的原貌,在体例术语与符号方面,我们基本上按原材料不变(有时受篇幅限制作了适当的删节),仅在最后附上几段“评注”,而这些评注也只是提出讨论,不一定完全正确,仅供参考。

§ 11.2 1991 年中国 5 岁以下儿童死亡抽样调查^{*)}

5 岁以下儿童死亡率是衡量一个国家是否真正发展的重要指标。1990 年 9 月 30 日在联合国召开的“世界儿童问题首脑会议”上通过一系列儿童战略目标,其中最重要和第一位的目标,是到 2000 年 5 岁以下儿童死亡率降低 1/3。1992 年 3 月国务院转发的“九十年代中国儿童发展规划纲要”提出的 10 项战略目标中,第一条也是到 2000 年 5 岁以下儿童死亡率降低 1/3。然而全国 5 岁以下儿童率基本还是一个空白。因此卫生部妇幼司决定首先在全国进行 1991 年中国 5 岁以下儿童死亡抽样调查,以搞清 1991 年中国 5 岁以下儿童死亡水平和死亡原因,为实现九十年代战略目标打下良好基础。在此基础上从 1992 年 1 月 1 日开始进行连续数年的监测和动态观察。

一、范围和对象

在全国 30 个省、自治区、直辖市范围内,抽取部分市、县的部分地区作为调查地区,将调查地区家庭中全部 0~4 岁儿童作为调查对象。

调查地区 1991 年孕满 28 周,娩出后有心跳、呼吸、脐带搏动、随意肌收缩四项生命指标之一均计为活产。1991 年调查地区 5 岁以下儿童死亡均填写儿童死亡报告卡。

*) 此项调查由卫生部妇幼司主持,首都儿科研究所具体负责研究与实施。作者参加了抽样设计工作。本节正文取自课题组研究总结报告。

二、抽样

1. 层的划分: 采用分层抽样技术, 将全国 30 个省、自治区、直辖市按地理位置(沿海、内地、边远)、经济发展程度及婴儿死亡率高低, 分为三大层, 其中四川省分为东西两部分。

A1(沿海): 北京、天津、上海、辽宁、山东、江苏、浙江、福建、广东。

A2(内地): 吉林、河北、河南、山西、安徽、湖北、湖南、广西、陕西、江西、海南、黑龙江、四川东部。

A3(边远): 内蒙古、宁夏、甘肃、新疆、青海、云南、贵州、西藏、四川西部。

每层内将市、县以“中国卫生状况分类”为基础, 每层分为六类: 即大城市、中小城市、一、二、三、四类县。沿海地区无四类县。全国 2377 个市县共分 17 小层。

2. 样本市县抽取: 抽取市县按以下原则进行:

(1) 每层抽取的市县数大致与该层市县总数成比例, 每层不少于 2 个。

(2) 抽取的样本在全国各省、市、自治区分布较为均匀。

(3) 每层抽取的样本市县加权平均婴儿死亡率接近该层加权平均婴儿死亡率。

(4) 适当考虑抽取县的条件。

按上述原则共抽取 81 个市县作为全国儿童死亡基础调查和监测网点。全国 81 个样本市县的分布, 样本加权平均婴儿死亡率及各层加权平均婴儿死亡率如表 11.1 所示。

表 11.1 儿童死亡监测市县分布及各层婴儿死亡率(IMR)

地区	大城市	中小城市	一类县	二类县	三类县	四类县	合计
沿海	样本数 5	5	6	3	3		22
	样本 IMR 13.2	18.7	23.5	18.6	21.6		
	层 IMR 13.4	18.5	22.3	21.8	19.5		
内地	样本数 7	9	6	10	9	2	38
	样本 IMR 16.4	19.3	27.6	33.4	34.2	50.2	
	层 IMR 18.4	21.8	28.2	33.3	33.3	59.0	
边远	样本数 2	2	2	4	8	3	21
	样本 IMR 22.2	53.7	49.1	47.6	54.9	98.8	
	层 IMR 24.8	46.0	46.1	40.7	56.2	91.1	
合计	9	16	14	17	20	5	81

3. 样本总量: 样本量 n 按下式计算:

$$n = \left[\frac{1.96}{d} \right]^2 pq, \quad p \text{ 是估计死亡率 } q = 1 - p, \\ d \text{ 为设计精度} = 2.5\%.$$

根据以上公式达到设计精度要求大层样本量为 200 万左右, 按大层及城乡分别计算样本总量应不少于 600 万人。

根据各层儿童死亡率的差异及监测条件等多种因素, 抽中市县样本人口数为:

大城市: 15~30 万,	二类县: 5 万,
中小城市: 8~15 万,	三类县: 4 万,
一类县: 5.5 万,	四类县: 3 万,

4. 区、乡抽样:

抽中市县的总人口均大大超过所需的样本量, 需进一步随机整群抽样。城市通常一个区即达到或超过规定样本量, 一般随机抽取一个城区(不包括郊区)。县中乡、镇数较多, 一个乡镇不能满足所需样本量, 一般要求抽样 2 个或 2 个以上的乡(镇), 抽样方法为系统抽样。

三、资料收集和质量控制(略)

四、资料分析方法

1. 分级加权:

(1) 由抽样地区调查的儿童死亡率(新生儿、婴儿、5 岁以下儿童), 计算每小层(17 小层)加权平均儿童死亡率。

(2) 由每小层的加权平均儿童死亡率, 计算六类地区(大城市、中小城市、一、二、三、四类县)加权平均儿童死亡率及三大层(沿海、内地、边远)加权平均儿童死亡率。

(3) 由六类地区加权平均儿童死亡率计算城市、农村及全国加权平均儿童死亡率。

2. 加权计算公式:

以上逐级加权均以 1982 年全国人口普查各区县人口数为基础, 计算公式为:

$$M_{ijk} = \frac{1991 \text{ 年调查地区 5 岁以下(或新生儿、婴儿)死亡数}}{1991 \text{ 年调查地区活产数}} \times 1000\%$$

$$M_{ij} = \sum \frac{N_{ijk}}{N_{ij}} M_{ijk}$$

$$M_i = \sum_{j=1}^3 \frac{N_{ij}}{N_j} M_{ij} \quad (\text{三大层, 城市, 全国加权按此计算})$$

M_{ijk} = 调查地区儿童死亡率; N_{ijk} = 调查市(县)总人口数

M_{ij} = 加权小层平均儿童死亡率; N_{ij} = 小层中调查市(县)总人口数

M_i = 分类(6类)加权平均儿童死亡率; N_i = 分类(6类)总人口数

3. 儿童死亡校正率的计算:

以乡、县、省三次核实补漏后, 省级上报卫生部的1991年抽中地区活产数和儿童死亡数作为基础调查数。以首都儿科研究所对三大层、六类地区1/3质量抽查发现的1991年活产和儿童死亡漏报数进行校正。

4. 校正率计算公式:

$$M_{校} = \frac{a \left[1 + \frac{a_{漏}}{a_{基}} \right]}{b \left[1 + \frac{b_{漏}}{b_{基}} \right]},$$

$M_{校}$ = 儿童校正死亡率;

a = 分类(分层)地区儿童死亡数;

b = 分类(分层)地区活产数;

$a_{基}$ = 抽查地区基础调查儿童死亡数;

$a_{漏}$ = 抽查地区漏报儿童死亡数;

$b_{基}$ = 抽查地区基础调查活产数;

$b_{漏}$ = 抽查地区漏报活产数。

5. 主要死因死亡率, 分类死因死亡率:

年龄别主要死因死亡率及年龄别分类死因死亡率均按上述方法进行加权和校正计算。死因构成以各省上报的死亡卡进行计算分析。

以年龄别校正死亡率作为全国、各层及各类地区的实际值。以年龄别校正死因死亡率作为各种死因死亡率的实际值。

评 注

1) 当确定要进行一项抽样调查时, 首要的问题是样本的抽取方法, 即抽样问题。最科学的抽样方法是概率抽样即随机抽样。但是在某些项目中, 严格的概率抽样在实施中往往有种种困难。例如本例的调查是一项全国性的调查, 必须首先对省(自治区、直辖市)进行抽样, 或直接对市、县进行抽样。但按概率抽样抽到的样本有时没有调查的条件。在本例中, 儿童死亡调查必须有周密的组织和较强的力量才能保证结果的准确。因此, 本例的抽样实际上(在层内抽样)是一种代表性的抽样。抽到的样本

是总体(全国)的一个缩影,各层样本婴儿死亡率(历史参考值)接近相应层的婴儿死亡率。在实际中,代表性抽样还是很有市场的。因为由此抽得的样本有相当好的对总体的代表性,调查结果也比较可靠、可信。但缺点是不能对调查结果给出精度的确切估计。

2) 分层是最常用的抽样技术之一。本例中的分层目的主要是为了提高精度。将全国市、县按地区以及城市与县分类,城市又按规模分,县又按《中国卫生情况分类》标准分为四类。这样可使调查结果的精度大为提高。至于分层(类)的标准是比较灵活的,中国卫生情况分类是按多项指标用聚类分析方法得到的。

3) 抽样调查误差包括抽样误差与非抽样误差两大类。在非抽样误差中,影响最大,也最难于控制与处理的是调查(测量)误差。在本例中,儿童死亡,特别是婴儿死亡在调查时误差较大,主要原因是遗漏。为了保证最后结果的可靠性,必须要求在调查过程中进行严格的质量控制,另外采取切实可行的措施进行校正。本例中的补漏即是一项重要措施,其作用绝对不可小估。

4) 资料处理在本例中即是总体目标量(例如5岁以下儿童死亡率)的估计。对于简单的线性估计而言,即是根据抽样方案(例如本例中的分层抽样或其他某些情形的不等概率抽样)给出不同的权数。当然如果设计是自加权的,则可省略这一步。

§ 11.3 全国办公自动化设备抽样调查^{*}

一、调查目的

为摸清我国微型计算机、复印机与传真机等办公自动化设备的拥有情况、用户使用状况及今后的市场需求。中国统计信息咨询服务中心在1991年7月至10月在全国范围进行一次抽样调查。在对调查资料汇总加工处理的基础上,中心会同有关专家及行业主管部门进行分析研究,于同年12月提出上述三种设备的市场分析报告,为管理、科研、生产、销售及维修部门的国内外客户提供必要的决策依据和信息。

二、抽样方法

采用分层二相随机抽样。

^{*} 本项目由中国统计信息咨询服务中心组织并实施。作者参加了其中设计与部分分析工作。本节正文取自该项目的抽样调查方案。

分层: 将全国按省、自治区、直辖市分成以下四层:

(1) 直辖市: 北京、天津、上海共三个市。直辖市作为自我代表层。

(2) 沿海省份: 包括河北、辽宁、江苏、浙江、福建、山东、广东、海南共 8 个省。

(3) 内地省份: 山西、吉林、黑龙江、安徽、江西、河南、湖北、湖南、四川共 9 个省。

(4) 边远省区: 内蒙古、广西、贵州、云南、西藏、陕西、甘肃、宁夏、青海、新疆共 10 个省区。

以上分层是根据各省、区、市国民经济水平及办公自动化设备的拥有情况来分的。根据多方面因素, 在全国抽取 17 个省、区、市作为调查点。再根据拥有情况, 除三个直辖市外, 确定沿海抽 5 个, 内地抽 6 个, 边远抽 3 个。在后三大层中利用简单随机抽样, 最后确定的省、区、市是:

(1) 直辖市: 北京、天津、上海。

(2) 沿海省份: 辽宁、江苏、福建、广东、浙江。

(3) 内地省份: 黑龙江、吉林、山西、江西、湖北、四川。

(4) 边远省区: 陕西、新疆、云南。

然后, 在以上每个抽中的省、自治区、直辖市中, 按二相抽样方法分别独立抽样。即先抽取一个较大的样本(第一相样本), 只调查各单位微型计算机、复印机、传真机的拥有及需求情况, 然后在拥有或有需求的样本单位中抽取一个较小的样本(第二相样本), 按调查问卷进行详细调查。这样既可以花有限的力量, 得到微型计算机、复印机、传真机拥有率的精确估计, 摸清这三种设备在我国未来三年内的需求量大小(第一相调查); 又可以集中人力、物力进行详细问卷调查(第二相调查)。

根据不同情况, 抽样工作按下述两种方法进行:

1. 直辖市抽样

直辖市抽样又分市区(包括郊区)和郊县, 市区作为一层, 郊县作为另一层。从郊县中随机抽取一个县作为调查点。

(1) 市区抽样

A. 第一相抽样:

首先将市区内所有单位分为 27 类(层)。根据各层估计的拥有率及各层单位总数, 确定各层样本量。各层样本的抽取可利用电话号码簿得到单位名单(大、中型工业企业应从其他途径如各地统计局工交处等得到, 大型工业企业还应得到其分厂的名单), 然后采用随机起点的等距抽样法

抽取第一相样本,进行第一相调查.

B. 第二相抽样:

根据第一相样本,并利用第一次调查结果,各层分别就微型计算机、复印机、传真机,将拥有及有需求的单位编号,各层得到三类单位清单(即微机拥有或有需求的单位清单、复印机拥有或有需求的单位清单、传真机拥有及有需求的单位清单),对各层每一类采取随机起点等概率系统抽样方式抽取样本,得到第二相样本(小样本),按调查问卷进行第二相调查.

(2) 县内抽样

A. 第一相抽样:

同市区第一相抽样.

B. 第二相抽样:

同市区第二相抽样.

2. 省、区抽样

每省、区抽两个城市和一个县. 在每个调查省(自治区)中将省会作为必抽城市,再从省会城市的郊县中利用简单随机法抽取1个县. 在其他市(地、州)中按简单随机抽样方法抽取一个作为另一个调查城市. 其中城市(包括省会)及县内抽样与直辖市市区及郊县的抽样相同.

三、确定样本量及分配

1. 第一相调查样本量的确定

根据实际情况,我们同时考虑全国和各大区的抽样精度,要求沿海地区(包括直辖市)抽样的绝对误差 $d_1 \leq 1.5\%$, 内地误差 $d_2 \leq 2\%$, 边远地区误差 $d_3 \leq 2\%$, 取置信度为 95%, 对简单随机抽样的比例估计, 在置信度 $1-\alpha$ 意义下, 若允许的最大绝对误差为 d , 则样本量由以下公式确定:

$$n = \frac{u^2 pq}{d^2},$$

其中: u 是标准正态分布的 α 双侧分位点, 当置信度为 95% 时, $\alpha = 0.05$, $u = 1.96$, 取 pq 最大值 0.25. 根据经验, 取设计效应为 $deff = 1.8$. 所需样本量如表 11.2 所示.

表 11.2

	绝对误差限	简单随机抽样所需样本量	实际需要样本量
沿海省市	1.5%	4268	7682
内地省份	2%	2401	4321
边远省区	2%	2401	4321

实际需要样本量是由简单随机抽样所需样本量扩大 $deff = 1.8$ 倍而得到的。

为了确保精度, 决定沿海地区抽取 8000 个单位, 内地省份抽取 4500 个单位, 边远省区抽取 4500 个单位, 全国共抽取 17,000 个单位, 这样全国的理论绝对误差为 1%。

2. 第二相调查样本量的确定

考虑到三种办公自动化设备的拥有率、样本的代表性及所要达到的调查精度, 第二相样本量分别为: 复印机 3200 (占第一相样本 18.8%)、微型计算机 2000 (占第一相样本 11.8%)、传真机 1000 (占第一相样本 5.9%)。

评 注

1) 随着市场经济的逐步建立, 人们工作条件与生活水平的不断提高以及各种产品的更新换代周期的缩短与剧烈的市场竞争, 市场调查愈来愈受到厂家及有关部门的关注, 市场调查的内容包括消费者的消费行为以及对商品(产品)的拥有与需求情况, 目的是及时地掌握产品的消费动向, 把握变化趋势。它对时效的要求较高, 而对设计要求不像其他抽样调查那样严格。对于许多市场调查, 例如消费者对产品的质量评价与满意程度等只需要抽一个不大的样本即可达到目的, 但是必要的抽样设计仍是不可缺少的, 特别是跨地域的较大规模的调查, 要考虑产品(现有的或潜在的消费者)的现实分布情况。目前在报刊上见到的某些市场调查并没经过科学的设计, 不一定能说明问题。例如在商店的家电柜台旁对顾客进行的家电需求调查就没有太大的意义, 因为他们只是一群特殊的顾客。

2) 市场调查一个最大的问题是抽样框的确定。本例中的调查对象是单位, 而单位一般都拥有电话, 因此利用电话号码簿作抽样框虽是一种不太严格(因为肯定会有遗漏)但却十分有效的方法。同时考虑到每一个单位不一定拥有所调查的三种设备, 或对其有所需求, 因此采用二相抽样的技术。先抽取一个大样本, 在第一相调查中只调查(估计)每种设备的拥有率及需求量, 然后在第一相样本中对拥有或有需求的单位抽取一个较小的第二相样本, 进行更详细调查, 这是二相抽样的一个实际应用。

3) 样本量的确定是每项抽样设计中的一个重要内容。样本量的确定取决于对精度的要求以及费用的限制, 对于要求达到的给定精度, 简单

随机抽样的样本量有比较简单的确定方法, 仅需对总体方差有大致估计即可。对于目标量为总体比例类型的量(在问卷调查中, 多数问题都以这种形式的目标量出现的), 总体方差可用其最大值 0.25 (当总体比例 $p=0.5$ 时) 代替, 以获得保守的估计。对于实际采用的复杂抽样, 要达到同样的精度, 需要乘上它的设计效应 $deff$ 。但在理论上仅对一阶整群抽样, 在已知群内相关的情形, 可用公式求得 $deff$ 值, 对其他复杂抽样, 则只有通过经验值进行估计了。在 § 11.5 的案例中, 就有对 $deff$ 的估计方法。

§ 11.4 全国粮食农药污染调查^{*}

根据国务院指示, 1984 年 3 月至 10 月由国家环境保护局与商业部、农牧渔业部共同组织了一次全国粮食受“六六六”与“滴滴涕”农药污染情况的大规模抽样调查。调查的目的是对全国各省、直辖市、自治区(除西藏、台湾外) 1983 年生产的主要粮食(小麦, 早、中、晚稻及玉米)中“六六六”和“滴滴涕”残留量的超标率、检出未超标率及未检出率与平均残留量作出全面而精确的估计。作为调查技术组的成员, 我们承担了制定粮食采样点分布方案(以下称抽样方案)及提出相应数据处理方法的工作。现将所用的抽样方案、目标量的估计与精度公式及其理论依据报告如下。

一、抽样方案

1 采样点的确定: 由于作为调查对象的粮食是一种散料, 因此在制定具体抽样方案前需确定基本抽样单元。我们选取乡级粮库作为基本抽样单元, 称为采样点。对每个被抽中的采样点, 根据粮食品种及不同的存贮方式, 按规定的方法, 采取有代表性的样品, 经充分混和后, 分取 1 kg 样品作为试样送检。这份试样完全作为相应采样点此种粮食的代表。例如若试样中“六六六”含量超标, 则相应采样点的该种粮食都按“六六六”含量超标计算。

2. 抽样方案的类型: 调查采用分层两级不等概率随机抽样法, 将 28 个省(市、自治区)作为层, 全部进行抽查。每层中第一级抽样(省抽采样县)采用与县的该种粮食产量成比例的概率无放回的抽取方法(详见 4)。第二级抽样(即采样县中抽采样点)则采用简单随机抽样, 每个采样县抽取数目相同(8 个)的采样点。采用这个方案的原因是: 样本代表性好, 实

* 本节正文引自冯士雍, 程翰生, 汪仁宫: 《全国粮食污染调查抽样方案的设计与数据处理方法》, 原载《应用概率统计》, 1985, 第 1 卷第 2 期, 155~160。

施方便,并有可能采用简单的数据处理方法。调查结果表明,所得估计量的精度较高。

3. 采样县与采样点数的确定: 样本量大小取决于调查精度和调查费用(工作量)之间的平衡。由于全国规模的粮食农药污染调查还是首次进行,缺少现成的污染程度及差异的有关资料,加之因时间紧迫不允许事先作试验性调查,因此只得从控制总工作量的前提下考虑样本量的大小,并对各层(省)作合理的分配。

从总的工作量考虑,各种粮食的采样点数以控制在5000~6000个为宜,即在所调查的粮食种类中平均每5万吨粮食取一个样。为保证全国及各省的调查精度,以及不使产量高的省工作量过大,我们采用各省每种粮食的采样县数(以及采样点数)与该省的这种粮食的产量的平方根成正比的原则确定。各省每种粮食按其产量所分配的采样县数见表11.3。

在制定方案时,尚缺各省(市、自治区)1983年分粮食种类的产量数据,因此在方案制定过程中都用1982年的相应数据代替。实际抽取的采样

表 11 3

实际产量(5万吨)	所需采样县数	实际产量(5万吨)	所需采样县数
0.3 以下	0	110 以上~132	11
0.3 以上~2	1	132 以上~156	12
2 以上~6	2	156 以上~182	13
6 以上~12	3	182 以上~210	14
12 以上~20	4	210 以上~240	15
20 以上~30	5	240 以上~272	16
30 以上~42	6	272 以上~306	17
42 以上~56	7	306 以上~342	18
56 以上~72	8	342 以上~380	19
72 以上~90	9	380 以上~	20
90 以上~110	10		

表 11 4

粮 食 种 类	实际采样县数	实际采样点数
早 稻	136	1088
中 稻	52	416
晚 稻	146	1168
小 麦	159	1272
玉 米	141	112
合 计	634	5072

样县及采样点数按粮食种类划分,见表 11.4.

方案还允许各省根据具体情况及需要,自设补充的采样县.但从这些采样点得到的数据在处理时不与按随机原则确定的数据混合,加上补充采样县,实际采样县总数为 679 个,采样点数为 5432 个.

4. 各省采样县的具体抽取方式: 省内抽取采样县是按各县该种粮食的产量大致成正比的不等概率随机抽样办法抽取的,具体抽取步骤是: 首先根据该省这种粮食的总产量在表 11.3 中查得所需采样县数,按各县的产量赋予每个县以与其产量成正比的代码个数(例如每 0.5 万吨一个代码),代码按全省各县级单位的自然顺序统一编号.若代码总数为 d ,则利用计算机产生 1 到 d 的(离散)均匀分布随机数,与所产生的随机数代码相应的县就作为抽中的采样县.直至所需的采样县数满足为止.

在抽取过程中,若一个县被抽取到两次或多于两次,则仍作为一个采样县处理.而以后面的随机数所代表的县依次递补.显然,实际采用的抽取方法是无放回的抽样方法.每次抽取时,每个当时还未被抽中的县被抽中为采样县的概率为该县产量对未被抽中县的总产量的比.即若令 Y_{hi} 为 h 省第 i 个县的产量, $Y_h = \sum_{i=1}^{N_h} Y_{hi}$ 为全省总产量,设前 $k-1$ 次抽中的采样县为 i_1, i_2, \dots, i_{k-1} , 则第 k 次抽中 i_k 县的概率为:

$$P\{\text{第 } k \text{ 次抽中 } i_k \text{ 县} \mid \text{前 } k-1 \text{ 次抽到 } i_1, i_2, \dots, i_{k-1} \text{ 县}\} \\ = \frac{Y_{h,i_k}}{Y_h - \sum_{i=1}^{k-1} Y_{h,i_i}} \quad (i_k \neq i_1, i_2, \dots, i_{k-1}). \quad (11.1)$$

二、对目标量的估计及其精度公式

调查数据的统计计算是根据粮食样品分析所得的“六六六”和“滴滴涕”残留量数据,结合 1983 年的实际产量,计算每种粮食按各采样县、各省及全国每种农药残留量的超标率、检出未超标率、未检出率和平均残留量的估计量和它们的精度.鉴于各种率的估计公式与精度公式对不同粮食、不同农药都是相同的,而对平均残留量也只须作少许变化就能采用同样的公式,因此下面仅以一种粮食一种农药的超标率为例,给出有关的计算公式.

1. 记号: 本节中涉及的各种主要记号的含义如下:

(1) 编号: 省编号 h , $h=1, 2, \dots, L$ ($L=28$); 县编号 i , h 省中实际县数为 N_h , 而采样县数为 n_h ; 县中采样点的编号为 j , $j=1, 2, \dots, 8$.

(2) 1983 年该种粮食的产量用 Y 表示, 1982 年产量用 Y' 表示. 特

别是, Y_{hij} 表示 h 省 i 县 j 点的 1983 年产量, $Y_{hi.}$ 为 h 省 i 县的产量, $Y_{h..}$ 为 h 省的总产量, $Y_{...}$ 为全国总产量. 若在 Y 上打上“ $''$ ”号, 则表示相应的 1982 年产量.

(3) 真实超标率记为 p , 相应的估计量记为 \hat{p} .

(4) $\lambda_{hij} = \begin{cases} 1, & \text{若 } h \text{ 省 } i \text{ 县 } j \text{ 点的粮食样品分析结果超标;} \\ 0, & \text{否则.} \end{cases}$

(注: 若令 λ_{hij} 为该点粮食样品分析结果的农药残留量, 则所有计算公式中的 p 即为平均残留量.)

2. 县超标率的估计: h 省 i 县的真实超标率:

$$\begin{aligned} p_{hi} &= \frac{h \text{ 省 } i \text{ 县 (该种粮食 1983 年) 产量中的超标部分}}{h \text{ 省 } i \text{ 县 (该种粮食 1983 年) 总产量}} \\ &= \frac{\sum_j \lambda_{hij} Y_{hij}}{\sum_j Y_{hij}} = \frac{\sum_j \lambda_{hij} Y_{hij}}{Y_{hi.}}. \end{aligned} \quad (11.2)$$

式中的求和是对县中的所有点进行的. 上式是一个比值. 由于在一个采样县中取的采样点数(8 个)相对于一般县中的总点数(乡的粮库数)比例较大, p_{hi} 可以用采样点的数值估计:

$$\hat{p}_{hi} = \frac{\sum_{j=1}^8 \lambda_{hij} Y_{hij}}{\sum_{j=1}^8 Y_{hij}}. \quad (11.3)$$

$E(\hat{p}_{hi})$ 与 p_{hi} 的偏差也甚小, 以下我们将这个偏差忽略不计, 即假定

$$E(\hat{p}_{hi}) \approx p_{hi}. \quad (11.4)$$

3. 省超标率的估计及方差计算: 省超标率可表成:

$$p_h = \sum_{i=1}^{N_h} p_{hi} \frac{Y_{hi.}}{Y_{h..}}. \quad (11.5)$$

根据各省中采样县的抽取方式, 即无放回的与产量有关的概率抽样(见公式 (11.1), 以下简称无放回抽样), 省超标粮食数 $Y_{h..} p_h$ 的估计可用 Murthy (1957) 的公式

$$Y_{h..} \hat{p}_h = \frac{\sum_{i=1}^{N_h} P(S|i) \hat{p}_{hi} Y_{hi.}}{P(S)},$$

从而

$$\hat{p}_h = \frac{\sum_{i=1}^{N_h} P(S|i) \hat{p}_{hi} Y_{hi.}}{P(S) Y_{h..}}, \quad (11.6)$$

其中 $P(S)$ 是 h 省中按无放回抽样抽到特定样本 S (大小为 n_h) 的无条件

概率, $P(S|i)$ 是在抽样中已知第一个抽到第 i 县而获得特定样本 S 的条件概率. 由于对固定的 i , $\sum_S P(S|i) = 1$ (其中求和是对所有大小为 n_h 的样本求的, 下同), 因此

$$\begin{aligned} E(\hat{p}_h) &= E[E_2(\hat{p}_h)] \approx E_1 \left[\frac{\sum_{i=1}^{n_h} P(S|i) p_{hi} Y_{hi}}{P(S) Y_{h..}} \right] \\ &= \sum_S \sum_{i=1}^{n_h} P(S|i) p_{hi} Y_{hi} / Y_{h..} \\ &= \sum_{i=1}^{n_h} p_{hi} Y_{hi} / Y_{h..} = p_h. \end{aligned} \quad (11.7)$$

从而 \hat{p}_h 是近似无偏的 (这里的近似仅是由于 (11.4) 式引起的).

在无放回抽样情形, \hat{p}_h 的方差估计为

$$v(\hat{p}_h) = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{n_h} [P(S|i, j) - P(S|i)P(S|j)] Y_{hi} Y_{hj} (\hat{p}_{hi} - \hat{p}_{hj})^2}{[P(S) Y_{h..}]^2}. \quad (11.8)$$

式中 $P(S|i, j)$ 为在前两个抽到第 i 县和第 j 县 (不考虑其次序) 情况下, 抽到特定样本 S 的条件概率. [严格地说, 为使 $v(\hat{p}_h)$ 是 $V(\hat{p}_h)$ 的无偏估计, (11.8) 式还应添加一项与第二级 (即是抽点) 抽样效应有关的小量, 参见 Cochran (1977) 第 11 章].

公式 (11.6) 与 (11.8) 的计算量非常大, 若用它们处理调查所得的所有数据有困难. 因此我们寻求替代办法.

将上述无放回抽样按与产量成正比的概率有放回抽样处理. 即在省抽县过程的 n_h 次抽样中, 每个县每次被抽到的概率都为 $Y_{hi}/Y_{h..}$, 则 p_h 可用 \hat{p}_{hi} 的算术平均数估计 (Cochran (1977) 第 11 章):

$$\hat{p}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{p}_{hi}, \quad (11.9)$$

它是 p_h 的无偏估计 (若不考虑 \hat{p}_{hi} 对 p_{hi} 的偏差). 而 \hat{p}_h 的方差的无偏估计为:

$$v(\hat{p}_h) = \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} (\hat{p}_{hi} - \hat{p}_h)^2, \quad (11.10)$$

作为有放回抽样的方差估计 (11.10) 比无放回抽样的方差估计 (11.8) 为大, 也即 (11.10) 式给出 \hat{p}_h 的方差估计的一个上限.

在应用上述公式时, 还有一个问题需要考虑: 在制定省抽县的方案时, 我们采用的是概率与每个县的 1982 年产量 Y'_{hi} 成比例, 而不是与

1983 年的实际产量 Y_{hi} 成比例。下面证明：只要假定 1983 年的污染程度与 1982 年的近似相等(粮食农药残留量主要与土壤残留的农药量及当年农药施用量有关，因此“六六六”与“滴滴涕”污染程度在它们完全停止使用前，相邻两年间的变化不会很大)，也即假定

$$p_{hi} \approx p'_{hi} \quad (i = 1, 2, \dots, N_h); \quad p_h \approx p'_h. \quad (11.11)$$

则上述结论，例如 \hat{p}_h 的无偏性亦近似成立。这是因为：

$$\begin{aligned} E(\hat{p}_h) &= E_1 E_2 \left(\frac{1}{n_h} \sum_{i=1}^{n_h} \hat{p}_{hi} \right) = \frac{1}{n_h} E_1 \left[\sum_{i=1}^{n_h} E_2(\hat{p}_{hi}) \right] \\ &\approx \frac{1}{n_h} \sum_{i=1}^{n_h} E_1(p_{hi}) = E_1(p_{h1}) \approx E_1(p'_{h1}) \\ &= \sum_{i=1}^{N_h} p'_{hi} \frac{Y'_{hi}}{Y'_{h..}} = p'_h \approx p_h. \end{aligned}$$

这里样本值与总体值用了相同的记号，实际期望号 E_1 下的 p_{hi} 和 p'_{hi} 为样本 S 的第 i 个样的值。

表 11.5 是两组数据分别按精确的无放回抽样公式(11.6)、(11.8)与按有放回近似公式(11.9)、(11.10)的比较。

表 11.5

	超标率的估计 \hat{p}_h		超标率的标准差 $\sqrt{V(\hat{p}_h)}$	
	按(11.6)式	按(11.9)式	按(11.8)式	按(11.10)式
样本 1 ($n_h=5$)	10.03%	10.14%	4.40%	4.89%
样本 2 ($n_h=7$)	6.26%	6.17%	2.66%	2.96%

从表 11.5 中可以看到近似公式(11.9)和(11.10)与精确公式(11.6)、(11.8)相差甚微。因此为计算方便起见，我们实际采用的是按有放回抽样的近似公式。

4. 全国超标率的估计与方差公式：

按分层抽样公式，从各省超标率的估计 \hat{p}_h 及其方差估计 $v(\hat{p}_h)$ 可得全国超标率

$$p = \sum_{h=1}^L p_h Y_{h..} / Y_{...} \quad (11.12)$$

的估计 \hat{p} 及其方差 $V(\hat{p})$ 的估计 $v(\hat{p})$ 如下：

$$\hat{p} = \sum_{h=1}^L W_h \hat{p}_h, \quad (11.13)$$

$$v(\hat{p}) = \sum_{h=1}^L W_h^2 v(\hat{p}_h), \quad (11.14)$$

其中层权

$$W_h = Y_{h..} / Y_{...} \quad (11.15)$$

是各省产量对全国总产量的比。只要 \hat{p}_h , $v(\hat{p}_h)$ 是 p_h , $V(\hat{p}_h)$ 的无偏估计, 则 \hat{p} , $v(\hat{p})$ 分别是 p , $V(\hat{p})$ 的无偏估计。

参 考 资 料

- [1] Cochran W G. Sampling Techniques, 3rd ed. John Wiley & Sons, 1977.
- [2] Murthy M N. Ordered and unordered estimators in sampling without replacement. *Sankhya*, 1957, 18: 379~390

评 注

1) 本例是一项专项调查, 目的比较单纯, 即估计全国 1983 年生产的主要粮食中“六六六”与“滴滴涕”残留量按卫生部规定的标准的超标率、检出未超标率与未检出率等, 但对估计精度要求较高。因此抽样设计必须严密, 并有与抽样方案匹配的目标量估计及其方差估计方法。本案例是严格遵循这个原则进行设计与分析的。

2) 抽样单元是每项调查必须首先明确的。对于像粮食这一类散料, 有许多种可能的选择。本例确定乡级粮库作为基本抽样单元, 因为它所储存的正是本乡或邻近乡生产的粮食, 没有从外运来的, 而且规模适中, 避免抽样的阶数过多。更为重要的是选中乡级粮库作单元, 顺理成章的就可以将行政系统作为抽样框使用, 从而简化了整个抽样过程。至于在粮库中样品采集的方法, 则采用粮食部门通用的随机采样法, 从库中的不同位置, 上下内外各层次中采集样品, 进行混合缩分制成试样, 按专门的化学定量分析方法得出“六六六”或“滴滴涕”的含量。这与社会经济调查中通常采用的问卷调查是完全不同的调查形式。对于这种专门调查, 测量(调查)误差较容易得到控制, 而且可以获得估计。

3) 本例的抽样方案采用分层二阶不等概率抽样。以省为层, 这是因为调查要求同时获得每个省的资料, 同时又便于调查的组织与实施。在层中的第一阶抽样, 即省中抽县用的是 Yates-Grundy 逐个抽取的与粮食产量基本成比例的概率抽样。采用不等概率抽样可较大程度地提高精度, 是在多阶抽样中的第一、二阶抽样中常用的方法。而 Yates-Grundy 方法又是对 $n > 2$ 情形中实施最方便的: 每次抽样都与所有未入样的单元(县)的大小(产量)成比例概率抽样。本例的另一个特点是: 第一阶抽样样本量 n (县数)的确定系按(省内)该种粮食总产量的平方根成比例。

这样既可保证每个省每种粮食都有一定数量的样本以保证精度,又能避免产量特别大的省份调查过多的县而造成浪费。这种折衷的方法是可取的。至于样本县中的第二阶抽样采用固定样本量($m=8$ 个乡级粮库)的简单随机抽样,是为了便于实施并简化数据处理。

4) 总体目标量若严格按照与抽样方案配套,则应采用 Murthy 估计量。而在 $n>2$ 时, Murthy 估计量计算公式 (11.6) 及其方差估计公式 (11.8) 都相当复杂,特别是后者。为此我们不得不进行简化,用放回 PPS 抽样公式代之。在本例中,我们经过实际计算比较两者的差异(即表 11.5),结果表明差异不大,近似程度是可以接受的。理论表明按放回 PPS 抽样的方差要比不放回的 PPS 抽样方差要大,表 11.5 的结果也证实了这一点。因此若仅就第一阶抽样,我们这样做所得到的实际方差的一个上限,是一个方差的较保守的估计。如果要校正这一点,还可以乘以估计的有限总体校正系数 $1 - f$, 其中 f 是第一阶抽样比的一个大致估计。不过由于我们将第二阶抽样引起的方差分量忽略掉了(理论与实际均表明这项数值相对于第一项要小得多),因此两者相抵,按本例计算的方差估计与实际数值应该相差不多。

§ 11.5 1987 年中国儿童情况抽样调查^{*}

根据中国政府与联合国儿童基金会 1985~1989 年合作项目计划,1987 年 7 月国家统计局会同有关单位组织内蒙古、黑龙江、浙江、山东、湖北、广东、四川、云南和宁夏等九省(区)进行了儿童情况的抽样调查。这次调查的标准时点是 1987 年 7 月 1 日零时。调查的工作时间是 1987 年 7 月 1 日至 7 月 31 日。本次调查的目的是掌握 0~14 岁儿童人数、儿童接受教育、生长发育、健康疾病、生存环境等情况。同时用这九省(区)的调查数据推算全国儿童的相应数据,为国家制定有关的方针政策、改善和加强儿童的卫生、保健、营养和教育等工作,加速培养四化建设人才提供科学依据。此外,由于中国儿童人数约占世界儿童人数的 1/6,因此这项调查的结果具有一定的世界意义。本文将详细介绍这项调查工作中的有关抽样设计,以及与抽样方案相适应的根据样本估计总体各目标量的公式及其相应的方差估计公式。最后就若干目标量的具体结果对上述

^{*} 本节正文摘自冯士雍与王思平《1987 年中国儿童情况抽样调查的抽样设计及数据处理模式》原载《中国儿童状况的调查与研究》,中国统计出版社,1990,32~46。

的设计与分析精度作出初步分析。

一、抽样设计

1. 抽样方案的类型

抽样设计应使调查具有充分的代表性, 保证一定的精度, 并尽可能节省调查的人力、物力与财力, 且便于组织管理。根据上述目的以及实际可能, 抽样调查方案采用分层二阶不等概率整群抽样, 在确定调查的每个省(自治区)中按城市及属于不同地形类型的县分层, 层内按与市、县的人口总数成比例的不等概率无放回方式抽取样本市(县)。在样本市(县)中按简单随机抽样方法抽取固定数量(10个)的样本点(基本上相当于村民委员会或居民委员会)。调查样本点(以下简称群)内所有0~14岁儿童。

2. 层的划分

儿童情况受所在地的经济文化水平和社会习俗等影响颇大, 同时考虑到调查组织管理的方便, 我们将各省、自治区作为大层。在每一大层中再按城市, 位于平原地区的农村县, 位于丘陵地区的农村县以及位于山区或高原地区的农村县共四种基本类型分为若干小层。若同一种类型中包含的县数过多, 则又按地理位置或行政区划细分为2~3个小层。其原则是每层抽2个市(县), 每省(自治区)中按市(县)的第一阶抽样比例大致为1/10(大省略低, 小省略高)。自然, 每省(自治区)不必一定包含上面所述的四种类别的小层。

按上述原则, 9省、自治区共有883个市、县共分为42个小层, 应抽样本市、县数为84个。本文以下部分的层皆是指上述的小层。

表 11.6 各省每种类型地区所包含的市县数及划分的层数

	城 市	平 原 县	丘 陵 县	山区或高原县	合 计
内 蒙 古	15(1)	15(1)	58(2)	—	88(4)
黑 龙 江	16(1)	21(1)	27(1)	14(1)	78(4)
浙 江	9(1)	23(1)	16(1)	28(1)	76(4)
山 东	18(1)	49(2)	19(1)	27(1)	113(5)
湖 北	13(1)	14(1)	14(1)	27(1)	68(4)
广 东	16(1)	25(1)	22(1)	46(2)	109(5)
四 川	14(1)	22(1)	59(2)	111(4)	206(8)
云 南	10(1)	—	8(1)	108(8)	126(5)
宁 夏	3(1)	8(1)	—	8(1)	19(3)
合 计	114(9)	177(9)	223(10)	369(14)	883(42)

3. 层内抽样本市(县)的方法

每层中按二阶抽样法, 第一阶段在层内抽市(县), 所用的方法是按与各市(县)的人口总数大致成比例的不等概率无放回方法抽取 $n=2$ 个样本市(县)。具体步骤是:

第一个样本市(县)的抽取是按照 $u_i = \frac{Z'_i}{Z'}$ 的概率随机抽取的, 这里的 Z'_i 是每个市(县)的 1984 年人口总数, 而 Z' 是当时该层所有市(县)的人口总数。设第 i 个市(县)被抽中。第二个样本市(县)是在剩下的市(县)中仍按与人口数 Z'_j 成比例的概率抽取, 从而实际是按 $u_j = \frac{Z'_j}{Z' - Z'_i}$ 的概率在剩下的市(县)中抽取的。

4. 样本市(县)中抽群(样本点)的方法

第二阶抽样是在样本市(县)中按简单随机抽样方法抽取 10 个样本点(群)。每个样本点基本上以居(村)民委员会为基础。为使抽样更有效率, 我们在抽样中对群的组成作了某些调整, 即按抽中的样本市(县)中各居(村)民委员会的名册及其相应的人口数将人口数相差过于悬殊的居(村)民委员会进行合并或分拆(分拆时以居(村)民小组为基础, 几个小组为一群), 使调整后的群所包含的人口数在同一个样本市(县)内大致相等。

按上述方案, 9 省(自治区)中共抽样本市(县)84 个, 样本点(群)840 个, 样本点中所包含的人口数(按 1984 年计)约为 77 万, 当时按儿童占总人口 $1/3$ 的比例计算, 所需调查的儿童数约为 25 万人。实际调查时所有样本点的总人口数(1987 年 7 月 1 日数字)为 811717 人, 其中调查儿童总数为 234659 人。

上述方案经联合国儿童基金会有关专家咨询后, 得到确认。在正式调查前, 各省还都组织过试调查。

二、各层目标量的估计及其方差估计

1. 关于调查目标量的简单说明

根据调查方案, 这次对儿童及其社会家庭环境因子的调查, 目标量共达 126 个之多。但从数据处理角度上说, 这些目标量大致可分为两类: 第一类是需给出有关总体总量的估计, 例如某一年龄组的儿童总数或某一类(例如独生子女)儿童总数; 另一类是关于两个这样总数的比, 例如学龄儿童(6~14 岁)在校率, 即是学龄儿童的在校人数与学龄儿童总人数之比值, 关于其他量, 例如平均值以及凡是在总人口(调查时的)中所占的某

类儿童的比例则可化成第一类目标量处理。

由于对不同目标量的总数估计与两个总数之比值的估计的处理方法都是同样的,因此本文只就这两种情况加以一般性讨论,给出数据处理的模式,主要包括目标量的估计量公式及估计量方差的估计公式,

2. 记号

本节只涉及层内数据的处理,故层编号省略。

i : 市或县编号,特别记样本市(县)的编号为 1, 2;

j : 群编号,特别记入样的群编号为 1, 2, \dots , m , 其中 m 一般等于 10;

$M_{i\cdot}$: i 市(县)中(经调整后的)群数;

$f_{i\cdot}$: i 市(县)中群的抽样比例,即 $m/M_{i\cdot}$;

Z_{ij} : i 市(县) j 群 1987 年 7 月 1 日时的人口总数;

$Z_{i\cdot}$: i 市(县) 1987 年 7 月 1 日的人口总数;

Z : 该层 1987 年 7 月 1 日时的人口总数;

Y , Y_i 及 y_{ij} 分别表示调查指标 y 的层总数, i 市(县)总数及 i 市(县) j 群总数,其中 Y 及 Y_i 的估计记为 \hat{Y} 及 \hat{Y}_i ;

X , X_i 及 x_{ij} 分别表示另一调查指标 x 的层总数, i 市(县)总数及 i 市(县) j 群总数,估计量的记号同上;

R , R_i 表示层及 i 市(县)中 y , x 指标总数之比值,即

$$R = \frac{Y}{X}, \quad R_i = \frac{Y_i}{X_i};$$

s^2 , $s_{\cdot\cdot}$ 分别表示 y_{ij} 或 x_{ij} 的样本方差或协方差,即

$$s_{y_i}^2 = \frac{1}{m-1} \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2,$$

$$s_{x_i}^2 = \frac{1}{m-1} \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2,$$

$$s_{y_i x_i} = \frac{1}{m-1} \sum_{j=1}^m (y_{ij} - \bar{y}_i)(x_{ij} - \bar{x}_i),$$

其中

$$\bar{y}_i = \frac{1}{m} \sum_{j=1}^m y_{ij}, \quad \bar{x}_i = \frac{1}{m} \sum_{j=1}^m x_{ij}.$$

3. 关于总数 Y (或 X) 的估计及方差估计

(1) 样本市(县)某目标量总数 Y_i 的估计

对 Y_i 的估计我们采用精度较高的对群人口的比估计,即令

$$\hat{Y}_i = \sum_{j=1}^m y_{ij} / \sum_{j=1}^m z_{ij} \quad (11.16)$$

作为 Y_i 与总人口数 Z_i 之比 $P_i = Y_i/Z_i$ 的估计。于是

$$\hat{P}_i = Z_i \hat{p}_i = Z_i \frac{\sum_{j=1}^m y_{ij}}{\sum_{j=1}^m z_{ij}}, \quad (11.17)$$

作为比估计, \hat{P}_i 是有偏的, 但偏倚并不大, 它的近似方差(从而也是近似均方误差)为

$$V(\hat{P}_i) = \frac{M_i^2(1-f_i)}{m} \frac{\sum_{j=1}^{M_i} (y_{ij} - \hat{P}_i z_{ij})^2}{M_i - 1}, \quad (11.18)$$

相应的估计量为

$$v(\hat{P}_i) = \frac{M_i^2(1-f_i)}{m(m-1)} \sum_{j=1}^m (y_{ij} - \hat{P}_i z_{ij})^2. \quad (11.19)$$

(2) 层中目标量总数 Y 的估计

在给出了层内两个样本市(县)目标量总数的估计 \hat{P}_1 、 \hat{P}_2 后, 根据我们给出的抽样方法, Y 的估计应用 Murthy 估计量, 在 $n=2$ 的情形, Murthy 估计量有以下较为简单的形式:

$$\hat{P} = \frac{1}{2} \frac{1}{u_1 u_2} \left[(1-u_2) \frac{\hat{P}_1}{u_1} + (1-u_1) \frac{\hat{P}_2}{u_2} \right]. \quad (11.20)$$

其中

$$u_1 = \frac{Z'_1}{Z'}, \quad u_2 = \frac{Z'_2}{Z'}. \quad (11.21)$$

Z'_1 、 Z'_2 、 Z' 是抽样时的相应人口数, 若忽略不计 \hat{P}_1 、 \hat{P}_2 的偏倚, \hat{P} 是无偏估计量。

根据二阶抽样的方差公式:

$$V(\hat{P}) = V_1[E_2(\hat{P})] + E_1[V_2(\hat{P})],$$

其中 E_1 、 V_1 是对第一阶抽样的期望与方差; E_2 、 V_2 是对给定的已抽得的一级单元(样本市、县)条件下抽样的期望与方差。通过计算得到

$$\begin{aligned} V(\hat{P}) = & \sum_i \sum_{i < j} \frac{u_i u_j (1-u_i-u_j)}{2-u_i-u_j} \left(\frac{Y_i}{u_i} - \frac{Y_j}{u_j} \right)^2 \\ & + \sum_i \sum_{i < j} \frac{1}{2-u_i-u_j} \left[\frac{u_j(1-u_j)}{u_i(1-u_j)} V(\hat{P}_i) \right. \\ & \left. + \frac{u_i(1-u_i)}{u_j(1-u_i)} V(\hat{P}_j) \right]. \end{aligned} \quad (11.22)$$

其中求和是对层内所有可能的市、县求的, 对 $V(\hat{P})$ 的估计, 我们采用以下无偏的估计量[参见 Brewer & Hanif(1989)]:

$$\begin{aligned}
 v(\hat{Y}) = & \frac{(1-u_1)(1-u_2)(1-u_1-u_2)}{(2-u_1-u_2)^2} \left(\frac{\hat{Y}_1}{u_1} - \frac{\hat{Y}_2}{u_2} \right)^2 \\
 & + \left(\frac{1-u_2}{2-u_1-u_2} \right)^2 \frac{v(\hat{Y}_1)}{u_1^2} + \left(\frac{1-u_1}{2-u_1-u_2} \right)^2 \frac{v(\hat{Y}_2)}{u_2^2} \\
 & - \frac{(1-u_1)(1-u_2)(1-u_1-u_2)}{(2-u_1-u_2)^2} \left(\frac{v(\hat{Y}_1)}{u_1^2} + \frac{v(\hat{Y}_2)}{u_2^2} \right).
 \end{aligned} \quad (11.23)$$

4. 两个目标量总数比值 R 的估计

(1) 估计量的形式

层中两个目标量总数 Y 与 X 的比值为

$$R = Y/X.$$

直接用上面的相应估计量 \hat{Y} 与 \hat{X} 的比值:

$$\hat{R} = \hat{Y}/\hat{X} \quad (11.24)$$

估计,而不必从样本市(县)中类似的比值出发.

(2) \hat{R} 的方差表示

由于(11.24)式中的 \hat{Y} 与 \hat{X} 都是通过一个复杂样本(二阶不等概率整群样本)估计,而 \hat{R} 又是非线性的估计形式,因此 \hat{R} 的方差是相当复杂的. 实际上,关于 \hat{R} 的方差,特别是方差的估计是这个项目理论中比较难于处理的问题.

利用 Taylor 级数展开,可获得一个随机变量任意函数的近似线性形式,例如采用上面的方法,可以得到

$$\begin{aligned}
 V(\hat{R}) \approx & \left(\frac{\partial \hat{R}}{\partial \hat{Y}} \right)^2 V(\hat{Y}) + \left(\frac{\partial \hat{R}}{\partial \hat{X}} \right)^2 V(\hat{X}) \\
 & + 2 \left(\frac{\partial \hat{R}}{\partial \hat{Y}} \right) \left(\frac{\partial \hat{R}}{\partial \hat{X}} \right) \text{Cov}(\hat{Y}, \hat{X}) \\
 = & \frac{1}{\hat{X}^2} V(\hat{Y}) + \frac{\hat{Y}^2}{\hat{X}^4} V(\hat{X}) - 2 \frac{\hat{Y}}{\hat{X}^3} \text{Cov}(\hat{Y}, \hat{X}) \\
 = & \hat{R}^2 \left[\frac{V(\hat{Y})}{\hat{Y}^2} + \frac{V(\hat{X})}{\hat{X}^2} - 2 \frac{\text{Cov}(\hat{Y}, \hat{X})}{\hat{Y} \hat{X}} \right]. \quad (11.25)
 \end{aligned}$$

(3) $V(\hat{R})$ 的估计

根据(11.25)式, $V(\hat{R})$ 的一个自然估计是

$$v(\hat{R}) = \hat{R}^2 \left[\frac{v(\hat{Y})}{\hat{Y}^2} + \frac{v(\hat{X})}{\hat{X}^2} - 2 \frac{\text{Cov}(\hat{Y}, \hat{X})}{\hat{Y} \hat{X}} \right]. \quad (11.26)$$

其中 $v(\hat{Y})$ 、 $v(\hat{X})$ 已由(11.23)式给出,而 $\text{Cov}(\hat{Y}, \hat{X})$ 是 $\text{Cov}(\hat{Y}, \hat{X})$ 的一个适当的估计. 因此现在的问题焦点在于给出 $\text{Cov}(\hat{Y}, \hat{X})$ 的形式.

为了表达简便起见,令

$$\begin{aligned}\alpha_1 &= \frac{1-u_2}{u_1(2-u_1-u_2)}, \\ \alpha_2 &= \frac{1-u_1}{u_2(2-u_1-u_2)}.\end{aligned}\quad (11.27)$$

这样, \hat{P} 、 \hat{X} 可分别表示成:

$$\begin{aligned}\hat{P} &= \alpha_1 \hat{P}_1 + \alpha_2 \hat{P}_2, \\ \hat{X} &= \alpha_1 \hat{X}_1 + \alpha_2 \hat{X}_2.\end{aligned}\quad (11.28)$$

由于 $E(\hat{P}) \approx Y$, $E(\hat{X}) \approx X$, 因此

$$\begin{aligned}\text{Cov}(\hat{P}, \hat{X}) &\approx E(\hat{P} - Y)(\hat{X} - X) \\ &= E[(\alpha_1 \hat{P}_1 + \alpha_2 \hat{P}_2 - Y)(\alpha_1 \hat{X}_1 + \alpha_2 \hat{X}_2 - X)] \\ &= E\{[\alpha_1(\hat{P}_1 - Y_1) + \alpha_2(\hat{P}_2 - Y_2) + (\alpha_1 Y_1 + \alpha_2 Y_2 - Y)] \\ &\quad \times [\alpha_1(\hat{X}_1 - X_1) + \alpha_2(\hat{X}_2 - X_2) + (\alpha_1 X_1 + \alpha_2 X_2 - X)]\} \\ &\approx \alpha_1^2 \text{Cov}(\hat{P}_1, \hat{X}_1) + \alpha_2^2 \text{Cov}(\hat{P}_2, \hat{X}_2) \\ &\quad + (\alpha_1 Y_1 + \alpha_2 Y_2 - Y)(\alpha_1 X_1 + \alpha_2 X_2 - X) \\ &\approx \alpha_1^2 \text{Cov}(\hat{P}_1, \hat{X}_1) + \alpha_2^2 \text{Cov}(\hat{P}_2, \hat{X}_2).\end{aligned}\quad (11.29)$$

为估计 $\text{Cov}(\hat{P}_i, \hat{X}_i)$ ($i = 1, 2$), 我们仿照相应的方差估计 $v(\hat{P}_i)$ [(11.19) 式], 用下式作为它的估计:

$$\text{Cov}(\hat{P}_i, \hat{X}_i) = \frac{M_i^2(1-f_i)}{m(m-1)} \cdot \sum_{j=1}^m (y_{ij} - \hat{P}_{i.} z_{ij})(x_{ij} - \hat{P}_{i.} z_{ij}), \quad (11.30)$$

其中

$$\hat{P}_{i.} = \frac{\sum_{j=1}^m y_{ij}}{\sum_{j=1}^m z_{ij}}, \quad \hat{P}_{i.} = \frac{\sum_{j=1}^m x_{ij}}{\sum_{j=1}^m z_{ij}} \quad (i = 1, 2). \quad (11.31)$$

将(11.30)代入(11.29)中的 $\text{Cov}(\hat{P}_i, \hat{X}_i)$, 即可得到 $\text{Cov}(\hat{P}, \hat{X})$ 的估计 $\text{Cov}(\hat{P}, \hat{X})$, 从而获得 $v(\hat{X})$.

三、各省及全国目标量的估计

1. 省目标量的估计

在给出了层内目标量的估计及其精度公式(方差估计公式)后, 利用分层抽样的有关公式就不难得到各省(自治区)相应目标量的估计.

设某省(自治区)由 L 层组成, 各层的层权为

$$W_h = \frac{h \text{ 层的人口数}}{\text{全省人口总数}} \quad (h=1, 2, \dots, L). \quad (11.32)$$

此时全省(自治区)某目标量总量 \tilde{Y} 的估计为

$$\hat{\tilde{Y}} = \sum_{h=1}^L \hat{Y}_h. \quad (11.33)$$

其中各层的 \hat{Y}_h 都由(11.20)式给出, $\hat{\tilde{Y}}$ 的方差估计是:

$$v(\hat{\tilde{Y}}) = \sum_{h=1}^L v(\hat{Y}_h). \quad (11.34)$$

$v(\hat{Y}_h)$ 由(11.23)式给出.

至于全省(自治区)两个目标量总量之比值 $\tilde{R} = \tilde{Y}/\tilde{X}$ 的估计为

$$\hat{\tilde{R}} = \sum_{h=1}^L W_h \hat{R}_h. \quad (11.35)$$

它的方差估计是:

$$v(\hat{\tilde{R}}) = \sum_{h=1}^L W_h^2 v(\hat{R}_h). \quad (11.36)$$

(11.35)式及(11.36)式中的 \hat{R}_h 与 $v(\hat{R}_h)$ 也由上节中相应的公式求得.

2. 全国目标量的估计

在此项调查中参加调查的9省(自治区)并不是从全国所有省、市、自治区中随机抽样得到的,但是在确定这些省(自治区)时是经过某种考虑的.首先是它们确有代表性,其次也考虑了调查工作开展的条件和方便.为获得全国相应目标量的估计及其精度,我们将全国所有省、市、自治区合成四种不同类型,而将调查的9省(自治区)分别归入适当的类型,从而可以看作是从中抽取的样本省.因而用分层抽样公式即可推得有关全国儿童的所有目标量的估计及相应的方差估计,具体公式从略.

四、部分目标量的估计结果及其精度

为表明本项目按上述抽样方案及数据处理模式获得的实际结果,表11.7列出了各调查省(自治区)及全国的以下7个目标量的估计;

1. 儿童总数 Y ;
2. 儿童在总人口中所占的比例 P ;
3. 独生子女儿童占儿童总数的比重 R_1 ;
4. 0~5岁儿童的入托率 R_2 ;
5. 6~14岁儿童的在校率 R_3 ;
6. 6~14岁儿童龋齿患率 R_4 ;
7. 婴儿死亡率 R_5 .

表 11.7 各省(自治区)及全国部分目标量的估计及精度

	内蒙古	黑龙江	浙江	山东	湖北	广东	四川	云南	宁夏	全国
\hat{P}	6 108 948	9 928 843	9 553 784	19 845 467	14 688 247	20 028 800	30 620 016	12 152 637	1 572 586	311 745 390
$S(\hat{P})$	234 928	110 690	469 806	763 388	552 146	986 275	690 454	234 269	47 697	8 880 153
$CV(\hat{P})$	0.0385	0.0111	0.0492	0.0385	0.0376	0.0492	0.0225	0.0193	0.0303	0.0124
\hat{P}	0.2932	0.2852	0.2332	0.2525	0.2916	0.3123	0.2942	0.3479	0.3662	0.2930
$S(\hat{P})$	0.0115	0.0033	0.0115	0.0097	0.0110	0.0154	0.0066	0.0067	0.0111	0.0036
$CV(\hat{P})$	0.0385	0.0111	0.0492	0.0385	0.0376	0.0492	0.0225	0.0193	0.0303	0.0124
\hat{E}_1	0.1386	0.2228	0.3304	0.2611	0.2022	0.1251	0.2183	0.1374	0.0984	0.2074
$S(\hat{E}_1)$	0.0152	0.0297	0.0230	0.0277	0.0189	0.0171	0.0164	0.0092	0.0154	0.0075
$CV(\hat{E}_1)$	0.0308	0.1335	0.0666	0.1056	0.0934	0.1366	0.0751	0.0672	0.1568	0.0362
\hat{E}_2	0.0976	0.1819	0.1983	0.2052	0.1722	0.0819	0.0930	0.1034	0.0723	0.1289
$S(\hat{E}_2)$	0.0130	0.0609	0.0270	0.0452	0.0225	0.0262	0.0144	0.0138	0.0188	0.0097
$CV(\hat{E}_2)$	0.1334	0.3348	0.1360	0.2262	0.1307	0.3204	0.1552	0.1335	0.2003	0.0754
\hat{E}_3	0.7665	0.7942	0.8197	0.7883	0.8115	0.7036	0.7948	0.7502	0.7173	0.7772
$S(\hat{E}_3)$	0.0471	0.0290	0.0578	0.0466	0.0570	0.0523	0.0374	0.0226	0.0395	0.0170
$CV(\hat{E}_3)$	0.0615	0.0366	0.0705	0.0591	0.0702	0.0743	0.0470	0.0301	0.0551	0.0219
\hat{E}_4	0.3261	0.4747	0.6047	0.4420	0.4342	0.6226	0.4563	0.5432	0.3447	0.4894
$S(\hat{E}_4)$	0.0406	0.0598	0.0510	0.0210	0.0382	0.0324	0.0456	0.0432	0.0231	0.0182
$CV(\hat{E}_4)$	0.1243	0.1261	0.0844	0.0475	0.0881	0.0841	0.0998	0.0795	0.0669	0.0371
\hat{E}_5	0.0446	0.0179	0.0274	0.0303	0.0461	0.0244	0.0354	0.0514	0.0353	0.0349
$S(\hat{E}_5)$	0.0140	0.0070	0.0087	0.0067	0.0033	0.0077	0.0067	0.0060	0.0060	0.0029
$CV(\hat{E}_5)$	0.3139	0.3911	0.3175	0.2175	0.0716	0.3156	0.1893	0.1107	0.1671	0.0831

表 11.7 对每个目标量 θ , 分别列出了估计量 $\hat{\theta}$, $\hat{\theta}$ 的标准差估计 $s(\hat{\theta}) = \sqrt{v(\hat{\theta})}$ 以及变异系数 $cv(\hat{\theta}) = s(\hat{\theta})/\hat{\theta}$.

根据表 11.7, 不难计算总体目标量 θ 真值的置信区间及估计量 $\hat{\theta}$ 的相对误差. 例如浙江省儿童在总人口中所占的比例 P 的 95% 的置信区间为

$$0.2332 \pm 1.96 \times 0.0115,$$

也即 (21.07%, 25.57%), 而 $\hat{P} = 23.32\%$ 在 95% 的置信水平下的相对误差为:

$$r(\hat{P}) = 1.96cv(\hat{P}) = 9.84\%.$$

对不同的项目, 调查的精度有所差别, 反映在估计量的变异系数上. 这是由于对不同调查项目, 有效样本量有较大差异. 例如在计算儿童在总人口中所占的比重时, 有效样本量是人口总数, 而在估计独生子女儿童在儿童中所占的比重时, 有效样本量应是儿童总数. 有效样本量大, 则估计精度较高. 此外, 对一些项目, 若调查误差大, 则由于这部分非抽样误差的影响, 也使估计量的标准差增大.

评 注

1) 这是我国国家统计局与联合国儿童基金会的合作项目, 1987 年的这次是第二次儿童情况调查. 此次调查在作设计前就已选定需要调查的 9 个省(自治区), 它们不是从全国省(市、自治区)中随机抽取的. 抽样设计实际上只对每个省(区)而言, 因此严格地说, 对总体目标量的估计只对调查省(区)有意义. 不过鉴于所调查的 9 个省(区)确实覆盖了全国不同类型的省(区), 我们采用事后分层方法, 把全国除台湾省以外的 29 个省(自治区、直辖市, 当时尚无海南省)分成四种不同类型, 并将 9 个调查省看作是从这四大类型省中抽出的随机样本, 再作全国目标量的估计. 这仅是一种不得已的办法. 不过在不准备对所有的省都进行调查的全国性项目, 这不是唯一的例子. 这其中既有主持单位种种特殊的考虑, 也有诸如调查组织, 甚至经费支持等实际因素, 此时抽样者唯一可以做的是在可能的条件下, 尽可能使调查的样本省有最好的代表性.

2) 为了提高精度, 每个调查省内按城市与县分层, 其中县按所处地理位置的地形状况分成三类. 这一点与 § 11.2 中国儿童 5 岁以下死亡抽样调查类似. 在我国, 处于平原、丘陵、山区或高原地区的经济文化水平差异甚大, 将它们分类作为不同层处理是合理的. 但在本例中并不将每个

省的县一律分成小层,而是按每类县的数目多少分成若干小层.例如四川省的丘陵县有 59 个,分成 2 个小层;山区或高原县有 111 个,分成 4 个小层.这样做的原因是:我们考虑抽样及以后数据处理的方便,将每个小层中抽取的市、县数 n 一律定为 2. 本例中在每个小层中采用的仍是 Yates-Grundy 逐个抽取法,这并不是一种严格的 π PS 抽样,但抽样方法比 Brewer 或 Durbin 方法简单,虽然数据处理稍为复杂些,但差别并不大.

3) 样本市县内的第二阶抽样是以居(村)民委员会为基础的整群抽样.为避免不同居(村)民委员会规模相差太大,事先经过适当调整,采用整群抽样是为了调查的便利,因为该项调查不仅使用通常的调查表(问卷)形式,也需有医生对每个儿童进行健康检查.样本太分散不利于实施.不过实际表明本例中的居(村)民委员会规模过大,似取居(村)民小组为群,更为合理,这样总样本量还可减少,且更能保证调查质量.不过以居(村)民小组为群要增加抽样的复杂性,即需要每个市、县具备以居(村)民小组为基本抽样单元的抽样框.在多数情形,这种条件不具备或需要专门去准备.当然一个可行的方法是在抽样中增加一阶抽样,在市、县中先抽取街道或居(村)委会,然后再抽居(村)民小组.不过这样又增加了数据处理的复杂性.

4) 本案例中的数据处理方法即第二段中的目标量估计及其方差估计是十分严格的,完全与抽样设计配套.作为总体总和估计公式(11.20)与方差估计公式(11.23)都是严格的.而对于两个目标量总数及它的比值 R 的估计方差,我们采用了第 9 章中的 Taylor 级数法.至于它的方差估计,关键在于 $\text{Cov}(\hat{P}, \hat{X})$, 本案例中利用(11.29)式将它化成 $\text{Cov}(\hat{P}_1, \hat{X}_1)$ 与 $\text{Cov}(\hat{P}_2, \hat{X}_2)$ 的估计.而后者的估计比较容易,这是本案例中的创新之处.

§ 11.6 北京地区专业技术人员现状抽样调查*

为摸清北京地区各种专业技术人员的基本情况,了解他们的愿望、要求及对当时改革中出现的许多问题的态度,由北京市科技干部局主持,在 1987 年组织了一次北京地区专业技术人员(含中小学教师)现状的抽样

* 本书正文引自马士雍、杨若勇:《北京地区专业技术人员现状抽样调查的抽样设计、数据处理方法和精度分析》,原载《应用概率统计》,1991,第 7 卷第 4 期,425~432.

调查。这次调查的另一个目的是配合全国就同一目的进行的抽样调查。鉴于北京具备其他省市所缺乏的某些特殊条件,北京地区的调查与全国范围内的调查在调查对象、抽样与问卷设计等方面都有所不同。但两者并不矛盾,经过适当技术处理,北京地区调查结果可以纳入全国调查结果中。

受主持单位委托,我们承担了此项调查的抽样设计,我们采用分层多阶不等概率抽样方法,对市属单位及中央在京单位各抽取250个基层单位,每个基层单位抽取10人,全部共计5000个专业技术人员进行了问卷调查。与此同时,我们给出了与设计相适应的从样本对总体各种目标量的估计及其精度的公式。对调查所得数据的处理结果表明,这次调查的精度完全达到了事先确定的设计要求。

一、抽样设计

1. 总体划分及抽样框的准备

此次调查的总体可分为两个子总体,即北京市属单位及中央在京单位的专业技术人员。经统计,1986年底北京市属单位专业技术人员数为398,140人;中央在京单位专业技术人员数为341,254人,合计共739,394人。

为抽样方便,将所有市属单位按系统归并为计委、经委等共14层,列出层内所属局一级单位名称及其专业技术人员数,形成完整的抽样框。同样,将中央在京单位按部、委、等部门列出共104个。如同市属单位一样,这些部门也可进一步细分,同时统计各部门专业技术人员数。

2. 样本量的确定与分配

我们根据对目标量估计的精度要求确定样本量。此次调查的目标量多数以比例形式出现。设 d 是在给定的置信水平 $1-\alpha$ 下,总体比例 P 的估计量 \hat{P} 的最大允许绝对误差,即 d 满足

$$Pr(|\hat{P}-P|\leq d)=1-\alpha, \quad (11.37)$$

则对简单随机抽样,当抽样比很小时,所需的样本量(按人计算):

$$m_0 = u_{\alpha}^2 P(1-P)/d^2. \quad (11.38)$$

其中 u_{α} 是标准正态分布的双侧 α 分位数。我们取 $1-\alpha=95\%$, $d=2\%$,此时 $u_{\alpha}=1.96\approx 2$,又 $P(1-P)$ 用它的最大可能值0.25代替,则根据(2.2)式有 $m_0=2500$,实际需要的样本量还需将它乘上设计效应(deff)。我们估计 $\text{deff}\approx 2$,于是实际需抽 $m=2500\times 2=5000$ 人。调查实际达到的精度及 deff 值的进一步估计在第三段中进行讨论。

由于实际调查采用派调查员面访形式,考虑到调查经费,人力以及效率,我们确定在每个被抽中的基层单位中调查 10 人。这样,全部共需调查 500 个基层单位。

所需调查的 500 个基层单位分配市属单位与中央单位各半。在考虑市属单位时,我们采用了最优分配。因为在市属单位的 14 层中,涉及区、县系统的两层,专业技术人员多集中在中小学及区、县属的医院与卫生院等,人员结构简单,情况相近,因而层内方差较小。此外,远郊区县由于交通不便,人均调查费用较大。于是将市属 14 层分成不属于区县的系统、远郊区县及城区与近郊区三大层。三大层的权按专业技术人员数分别为 $W_1=61\%$ 、 $W_2=15\%$ 和 $W_3=24\%$ 。假定有关区县的两层的层内方差是其他系统层内方差的 $1/2$, 即 $\frac{1}{2}S_1^2=S_2^2=S_3^2$, 远郊区县的单位人员调查费用是其他两层的 2 倍, 即 $c_2=2c_1=2c_3$, 则根据分层抽样中的最优分配, 每大层样本量(为方便起见, 以下以基层单位数计算):

$$n_h \propto n W_h S_h / \sqrt{c_h} \quad (h=1, 2, 3). \quad (11.39)$$

其中 $n=250$, 由此可计算得 $n_1=178$, $n_2=22$, $n_3=50$ 。其中第一大层也即非区县的 12 个系统(层)共需抽 178 个基层单位, 这些单位按各层大小即专业技术人员数成比例的原则分配。

至于中央单位需抽的 250 个基层单位, 由于具体的抽样不是按分层进行的, 故不需事先进行分配。但根据下面第 3 段中所述的方法, 在 104 个部门中实际抽中的基层单位数基本上也与各部门中的专业技术人员数成比例。

3. 具体抽样程序

由于两个子总体抽样框形式及具体条件不尽相同, 因此采用的抽样方法也有所不同。

对市属单位, 采用分层多阶抽样。其中在有关区、县两层内, 又采用按多种方式进一步分层技术: 既按不同的区、县分, 又按单位的性质分。具体地说, 城区近郊区层与远郊区县层又稍有差别。对于后者, 为使样本单位进一步集中, 先将所属 10 个区县按经济发展水平分为两小层, 然后按简单随机抽样分别在两小层中共抽取 5 个区县进行调查。对城区近郊区层, 则将每个区作为一小层。以上各区县的所有单位都按学校、医院(或卫生院)及其他单位分为三类(也作为小层处理)。在以上分层(类)中都按比例分配的原则分配所需调查的基层单位数。至于各区县每个小层内

的抽样则是按简单随机抽样抽取基层单位, 在每个被抽中的基层单位中按简单随机抽样或等距抽样抽取 10 人的方法。

在区、县以外的其他 12 层内, 我们都采用三阶抽样, 即层内抽局级单位, 局级单位内抽基层单位, 基层单位内抽人的方法。

第一阶抽样即层内抽局级单位的方法是按(层级)单位大小成比例的放回不等概率(PPS)抽样。具体方法是: 设该层共有 N 个局级单位, 第 i 个局级单位有专业技术人员 M_i 人。令

$$M_0 = \sum_{i=1}^N M_i, \quad z_i = M_i / M_0,$$

又设分配给该层的样本量为 n 个基层单位, 则需独立地做 n 次放回随机抽样, 每次第 i 个局级单位的入样概率为 z_i 。记 n_i 为在这 n 次抽样中, 第 i 个局级单位入样的次数, 此数即为该局级单位内需调查的基层单位数, n_i 可能为 0。

第二阶抽样即是在第一阶抽样中被抽中的局级单位中抽取所需要数量的基层单位。其方法仍是 PPS 抽样(但不放回, 在抽样中重复抽中的不计, 直到抽到不同的且满足要求数量的基层单位为止)。对于其中基层单位大小相差不多的局级单位也采用简单随机抽样。

第三阶抽样是在每个被抽中的基层单位中抽取 10 人进行实际调查。方法是按该单位专业技术人员的名册用简单随机或等距抽样方法抽取。

中央在京单位的抽样也采用上述市属单位层内采用的三阶抽样法。即在全部 104 个部门中用 PPS 抽样部门, 独立重复 250 次。每个部门被抽中的次数即为该部门中所需抽的基层单位数。结果有 76 个部门被抽中。至于在这些部门中抽基层单位以及在基层单位中抽人的方法与市属单位层内第二、三阶抽样完全相同。

二、数据处理公式

1. 目标量的分类

根据问卷, 此次调查共有 107 个问题, 647 个选择项, 833 个调查指标或需要估计的总体目标量。这些目标量从其形式可分成以下四类:

(1) 总体或子总体的总值 Y , 即某个指标 y 的总体(或在统计范围内)的总和, 例如北京地区专业技术人员家庭居住总面积等;

(2) 总体或子总体的平均数 \bar{Y} 。例如平均月收入等;

(3) 总体比例。按某种类别或准则分类的专业技术人员在全专业技术人员中所占的比例。例如 1978 年以来出过国或去过港澳地区人员

在全体专业技术人员中所占的比例等;

(4) 两个总体总量或平均数的比值 $R = Y/X = Y/\bar{x}$, 其中 x 是另一个指标. 例如由于专业不对口而造成工作时间内任务不足的人员的比例, 即是对问题 308 作该项选择(应填 1 者)的总人数 Y 与对该问题所有选择非零的(即认为他的工作时间内任务不足者)总人数 X 的比值. 这里 Y 与 X 都需要估计.

在以上四类目标量中, 前三类都可归结为总量的估计. 故以下我们只需对 Y 、 R 这两类本质不同的目标量进行讨论.

2. 记号

为表达方便, 重新规定各记号的意义如下:

以 Y 、 X 或加上适当的下标记指标值 y, x 在一定范围内的总和, 记 $R = Y/X$ 为它们的比值, 而 \hat{Y} 、 \hat{X} 、 \hat{R} 为相应的估计量; $v(\cdot)$ 与 $s(\cdot) = \sqrt{v(\cdot)}$ 分别表示估计量的方差和标准差的估计.

以 h 为层的编号(在不会引起混淆的情形也常被省略); i 为中央部门或市属局级单位的编号(为简便起见不再区分总体中的和样本中的), 在样本中也表示基层单位编号; j 表示被调查者的编号.

M_h 表示 h 层内专业技术人员数, $M_0 = \sum_h M_h$ 为所考虑范围内专业技术人员总数, $W_h = M_h/M_0$ 为层权, $z_{hi} = M_{hi}/M_h$ 为 h 层内 PPS 抽样中第 i 个(局级)单位每次抽样中的入样概率, m_{hi} 表示 h 层第 i 个(样本基层)单位中回收的有效问卷数.

3. 中央单位的数据处理公式

对每个被调查的基层单位, 计算

$$\hat{Y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}, \quad \hat{X}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}, \quad (11.40)$$

$$\hat{Y}_i = M_i \hat{\bar{y}}_i, \quad \hat{X}_i = M_i \hat{\bar{x}}_i. \quad (11.41)$$

其中 y_{ij} 、 x_{ij} 是该单位中第 j 个被调查者对问题回答的指标值, M_i 是该单位所在部门专业技术人员数, \hat{Y}_i 与 \hat{X}_i 是无偏的.

由于对部门的抽样(决定抽哪些部门以及每个部门中抽几个基层单位)是按有放回的 PPS 抽样决定的, 故对总体总和估计应采用下述的 Hansen-Hurwitz 估计量(见参考资料[1]):

$$\hat{Y} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i / z_i = \frac{M_0}{n} \sum_{i=1}^n \hat{\bar{y}}_i, \quad \hat{X} = \frac{1}{n} \sum_{i=1}^n \hat{X}_i / z_i = \frac{M_0}{n} \sum_{i=1}^n \hat{\bar{x}}_i. \quad (11.42)$$

这里 $z_i = M_i/M_0$, M_0 是中央在京单位专业技术人员的总数, 而 $n=250$. \hat{Y} 、 \hat{X} 也是无偏的.

对于比值型目标量 R 的估计, 我们用

$$\hat{R} = \hat{Y} / \hat{X}. \quad (11.43)$$

至于 \hat{Y} 的方差, 若忽略第三阶抽样的误差, 按二阶抽样计算, 有

$$V(\hat{Y}) = \frac{1}{n} \sum_{i=1}^N z_i \left(\frac{Y_i}{z_i} - Y \right)^2 + \frac{1}{n} \sum_{i=1}^N \frac{M_i^2 (1 - f_{2i}) S_{2i}^2}{m_i z_i}. \quad (11.44)$$

其中 f_{2i} 是第 i 部门中二阶抽样比例, S_{2i}^2 是第 i 部门内总体方差, N 是部门总数, 即 104. (11.44) 式适用于第一阶抽样为放回 PPS 抽样, 而第二阶抽样中每个样本单元 (相当于基层单位) 也必须放回总体的情形. 实际抽样对基层单位和人员都是不放回的. 若将后二阶抽样合在一起作为简单随机抽样处理, 则实际方差应比 (11.44) 稍小 (参见参考资料 [3]), 也即应为

$$V(\hat{Y}) = \frac{1}{n} \sum_{i=1}^N z_i \left(\frac{Y_i}{z_i} - Y \right)^2 + \frac{1}{n} \sum_{i=1}^N \frac{M_i^2 (1 - f_{2i}) S_{2i}^2}{m_i z_i} - (n-1) \sum_{i=1}^N \frac{M_i S_{2i}^2}{n}. \quad (11.45)$$

由于第一阶抽样是放回的 PPS 抽样, 从而 \hat{Y}_i 是相互独立的, 且如前已指出的, \hat{Y}_i 是无偏的, 因此据 (见参考资料 [1]):

$$v(\hat{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{\hat{Y}_i}{z_i} - \hat{Y} \right)^2, \quad (11.46)$$

是 $V(\hat{Y})$ 的一个无偏估计量. 同样对 \hat{X} 有

$$v(\hat{X}) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\frac{\hat{X}_i}{z_i} - \hat{X} \right)^2. \quad (11.47)$$

为推导 \hat{R} 的方差及其估计, 我们利用 $\hat{R} = \hat{Y} / \hat{X}$ 的 Taylor 展开, 取其线性项, 求其方差可得:

$$V(\hat{R}) \approx \hat{R}^2 \left[\frac{V(\hat{Y})}{\hat{Y}^2} + \frac{V(\hat{X})}{\hat{X}^2} - 2 \frac{\text{Cov}(\hat{Y}, \hat{X})}{\hat{Y} \hat{X}} \right]. \quad (11.48)$$

(11.46) 与 (11.47) 式已给出 $V(\hat{Y})$ 及 $V(\hat{X})$ 的估计, 为获得协方差 $\text{Cov}(\hat{Y}, \hat{X})$ 的估计, 令

$$u_{ij} = x_{ij} + y_{ij}, \quad (11.49)$$

将 u 作为另一个指标. U_i 、 U 的估计以及 \hat{U} 的方差估计 $v(\hat{U})$, 都可用与 \hat{Y}_i 、 \hat{Y} 、 $v(\hat{Y})$ 相类似的公式计算. 另一方面, 由 $\hat{U} = \hat{Y} + \hat{X}$ 知

$$\text{Cov}(\hat{Y}, \hat{X}) = \frac{1}{2} [V(\hat{U}) - V(\hat{Y}) - V(\hat{X})]. \quad (11.50)$$

于是 $\text{Cov}(\hat{P}, \hat{X})$ 可用下式估计:

$$\text{Cov}(\hat{P}, \hat{X}) = \frac{1}{2} [v(\hat{O}) - v(\hat{P}) - v(\hat{X})]. \quad (11.51)$$

综合(11.48)与(11.51)式, $V(\hat{R})$ 的估计为

$$v(\hat{R}) = \hat{R}^2 \left[\frac{v(\hat{P})}{\hat{P}^2} + \frac{v(\hat{X})}{\hat{X}^2} - \frac{v(\hat{O}) - v(\hat{P}) - v(\hat{X})}{\hat{P}\hat{X}} \right]. \quad (11.52)$$

4. 市属单位的数据处理公式

根据抽样方法, 市属单位中有关区、县两层内采用多种方式的仔细分层, 而小层内采用二阶简单随机抽样. 为简化起见, 我们采用分层随机抽样公式处理. 设每层内又分为 L 小层, h 小层的专业技术人员数为 M_h , 总人数为 M_0 , $W_h = M_h/M_0$, 则

$$\hat{P} = M_0 \hat{\bar{Y}} = M_0 \sum_{h=1}^L W_h \bar{y}_h = \sum_{h=1}^L \hat{P}_h, \quad (11.53)$$

其中

$$\bar{y}_h = \frac{1}{m_h} \sum_{j=1}^{m_h} y_{hj} \quad (11.54)$$

是 h 小层 y 指标的样本平均数, \hat{P} 的方差估计(由于抽样比 f_h 很小, 忽略不计)为

$$v(\hat{P}) = \sum_{h=1}^L \frac{M_h^2}{m_h(m_h-1)} \sum_{j=1}^{m_h} (y_{hj} - \bar{y}_h)^2. \quad (11.55)$$

对于小层内比值型目标量 R_h 的估计则为

$$\hat{R}_h = \hat{P}_h / \hat{X}_h, \quad (11.56)$$

它的方差估计为

$$v(\hat{R}_h) = \frac{1}{m_h(m_h-1) \bar{x}_h^2} \sum_{j=1}^{m_h} (y_{hj} - \hat{R}_h x_{hj})^2, \quad (11.57)$$

其中

$$\bar{x}_h = \frac{1}{m_h} \sum_{j=1}^{m_h} x_{hj}, \quad (11.58)$$

于是

$$\hat{R} = \sum_{h=1}^L W_h \hat{R}_h, \quad (11.59)$$

而

$$v(\hat{R}) = \sum_{h=1}^L W_h^2 v(\hat{R}_h). \quad (11.60)$$

至于除区、县以外的其余市属各系统 12 层, 由于层内抽样方法与对中央在京单位抽样完全相同, 因此估计量及其方差估计也完全与上面第 3 段中的相应公式相同, 只要加上层的编号即可. 而为得到市属所有 14

层的汇总结果,则再次根据分层抽样的公式,即可得到。

5. 全北京地区目标量的估计与方差估计

全北京地区目标量的估计是将中央单位与市属单位作为两大层考虑。将上面第3段与第4段中所得的结果也用分层抽样有关公式汇总即可得到全北京地区目标量的估计及其方差估计。

三、部分调查结果的实际精度及抽样的设计效应

1. 样本回收情况及质量

实际调查采用派调查员面访的形式,所有调查员经过短期培训,在正式调查前又进行了试调查。因此调查质量较高,问卷回收率达到100%,其中有效问卷率为99.94%。故可以排除不回答引起的非抽样误差。在数据录入前后的各个环节都进行了严格的质量控制。

2. 目标量估计的实际精度

在一、2段中我们规定了调查的设计精度是在置信水平95%下,关于比例型估计量的绝对误差不超过2%。根据前面给出的方差估计公式,对每个目标量进行计算,即可获得对每个目标量估计的实际精度的估计。上段中的精度都是以 $v(\cdot)$ 形式给出的,若换算成给定置信水平95%的最大绝对误差 d^* ,有以下关系:

$$d^* = u_{\alpha} \sqrt{v(\cdot)} = 1.96s(\cdot). \quad (11.61)$$

此外,估计量的精度也常用最大相对误差 r 或变异系数 cv 表示。 r 也是对一定置信水平意义下而言的。例如对 Y 的估计量 \hat{Y} , r 满足

$$Pr\left(\left|\frac{\hat{Y}-Y}{Y}\right| \leq r\right) = 1-\alpha, \quad (11.62)$$

它与 $s(\cdot)$ 及 $cv(\cdot)$ 之间有关系

$$r = u_{\alpha}s(\hat{Y})/\hat{Y} = u_{\alpha}cv(\hat{Y}). \quad (11.63)$$

表11.8列出了此次调查全部833个以比例或比值形式的总体目标量估计的标准差的分布情况。

如果换算成实际达到的最大绝对误差 $d^* \approx 2s$,则可看到97.3%的最大绝对误差小于等于设计精度2%,这个比例比规定的置信水平95%高,可见调查完全达到了事先要求的精度。

3. 设计效应(deff)

显然,精度必须在相同样本量下进行比较才有实际意义。Kish(见参考资料[2],引进称为设计效应(deff)的量来表示一个复杂抽样设计的效率:

表 11.8 有关比例或比值型目标量估计的标准差分布

标 准 差 范 围	频 数	频 率
$s \leq 0.2\%$	156	16.8%
$0.2\% < s \leq 0.5\%$	297	35.7%
$0.5\% < s \leq 1.0\%$	358	43.0%
$1.0\% < s \leq 1.5\%$	21	2.5%
$1.5\% < s \leq 1.8\%$	2	0.2%
合 计	833	100%

$$\text{deff} = \frac{\text{按照复杂抽样设计估计量的方差}}{\text{按简单随机抽样同样本量时估计量的方差}} \quad (11.64)$$

例如对一个比例型估计量 \hat{P} , 对简单随机抽样(SRS), 当总样本量为 m 时, 它的方差估计(当忽略有限总体校正系数时)为:

$$v_{\text{SRS}}(\hat{P}) \approx \hat{P}(1 - \hat{P})/m, \quad (11.65)$$

于是此次调查的设计效应的估计为:

$$\text{deff} = \frac{5000v(\hat{P})}{\hat{P}(1 - \hat{P})} \quad (11.66)$$

若将所有可以化成比例形式的总量估计都化成 $\hat{P} = \hat{Y}/M$ 的形式(其中 M 为总体或子总体的大小), 对比值 \hat{P} 也借用(11.66)式, 即可计算具体的 deff 值。表 11.9 是问卷中某个问题各选择项目标量的估计值, 标准差、变异系数与 deff 的估计值。

表 11.9 专业技术人员任务量不足情况及其原因分析(问题 308)

问 题 选 择	$\hat{P}(\%)$	$s(\hat{P})\%$	$cv(\hat{P})(\%)$	deff
0. 无此情况(任务饱满)	68.120	0.9	1.3	1.86
1. 专业不对口	2.279	0.2	8.8	0.90
2. 无合适工作	1.700	0.2	11.8	1.20
3. 分配不当	3.124	0.3	9.6	1.49
4. 工作条件不具备	6.434	0.4	6.2	1.33
5. 健康原因	0.912	0.1	11.0	0.55
6. 没有任务	7.282	0.4	5.5	1.18
7. 人多事少	4.062	0.3	7.4	1.15
8. 领导不分配工作	2.155	0.2	9.8	0.95
9. 其他原因	3.933	0.3	7.6	1.19

对随机选择的 15 个问题共 113 个选择项所作的统计表明: 变异系数有 46.02% 不超过 5%, 69.03% 不超过 7.5%, 81.4% 不超过 10%; deff 有 66.37% 不超过 1.5, 83.19% 不超过 2, 其平均值为 1.56. 在同类设计中, deff 的这个值相当小, 这表明此次抽样设计的效率较高.

参 考 资 料

- [1] Cochran W G. Sampling Techniques, 3rd ed. Wiley & Sons, 1977.
- [2] Kish L. Survey Sampling, Wiley & Sons, 1965.
- [3] Sukhatme P V. Sampling Theory of Surveys, With Applications, Iowa State College Press, 1954.
- [4] Feng Shi-yong, Wang Si-ping. The Design and Data Processing of the Sampling Survey of Children's Situation in China 1987, Acta Mathematica Applicatae Sinica, 1990, Vol. 6(4): 351~360.

评 注

1) 本案例是对所有在京的北京市属与中央所属单位近 74 万专业技术人员的现状进行的调查. 采用典型的问卷调查形式, 共调查了 107 个问题, 包括 647 个选择项. 与一般问卷调查类似, 除了极少数问题需要作出定量回答外, 其他问题仅需圈填问卷中所列的选择项. 因此总体目标量即是圈填每项选择项的比例, 即 P . 对这样的问题样本量较易确定, 首先按简单随机抽样估计精度公式, 对于 $P=0.5$, 即 $PQ=0.25$ 这一最保守(总体方差最大)的情形, 对给定的对 P 估计量绝对误差限 d 及相应的置信度, 即可确定简单随机抽样的样本量 n' . 若取置信度为 95%, $d=1\%$, 则相应的 $n'\approx 10000$; 若取 $d=2\%$, 则 $n'\approx 2500$. 一般的 d 不宜取得太小, 也不宜取得太大, 否则, 不是需要的样本量太大, 就是精度不够, 结果不可靠. 通常取 d 在 $1\%\sim 3\%$ 范围内. 实际样本量 $n=n'/\text{deff}$, 即还要乘上设计效应, 而设计效应可根据对类似调查的经验而定. 例如在本案例中事先估计 $\text{deff}=2$, 而根据按实际调查结果的估计量的方差估计的具体结果与同样样本量的简单随机抽样比较(如本案例第三段中所述)即可获得 deff 的估计. 虽然就每个指标而言, 这种估计是不同的, 但可以根据它的分布确定, 以作今后类似设计时的参考. 在本例中, 对随机选择(因没有对所有项都进行 deff 计算)的 113 个选择项的统计, 有 66.37% 的 deff 不超过 1.5, 83.19% 的项不超过 2. 因此原先估计的 2 还是比较

合适的,如果要求不太高,取 d_{eff} 为 1.8 左右也就可以了。

2) 在设计时遇到的一个重要问题是抽样框的编制,选用合理且方便的抽样框是实施抽样的前提。基于每个专业技术人员都属于一个独立的基层单位(不独立的单位或兼职的单位不算),而每个基层单位又有所隶属的上级单位或主管单位,因此我们将基层单位作为基本抽样单元,在每个被抽中的基层单位中抽取被调查的专业技术人员,人数固定为 10 人,这样便于操作,效率也较高。当然 10 人是不是最佳选择与每调查一个单位需耗费的人力与时间有关,这里仅是直观上觉得比较合理的数值。

3) 中央在京单位与北京市属单位是作为两个子总体独立抽样的,各抽 250 个基层单位。其中中央单位的隶属系统比较简单,按部门即可获得其所属的所有基层单位的名册及专业技术人员数,因此只需用二阶抽样即可,这比采用更高阶的抽样效率高。由于各部门中的专业技术人员数相差很大,因此宜用不等概率抽样,本例中采用放回 PPS 抽样,这样不仅实施方便,而且数据处理也简单。按所分配的 250 个基层单位的样本量,对全部 104 个部门独立进行 250 次抽样,以每个部门被抽中的次数 n_i 作为该部门需抽的基层单位数,按简单随机抽样或随机起点的系统抽样在该部门中抽取。这是一种比较巧妙的方法。结果表明,按这种抽样与将部门作层的比例分配的分层抽样的结果非常接近,差别仅是对那些规模较小的部门不一定能保证被抽中而已。

4) 市属单位的隶属关系比较复杂,上、下级层次较多,专业技术人员集中与散布的情况差别很大,因此我们仔细地进行了分层,首先是分不属于区、县管理的 12 个系统作为一大层,远郊区县与市区、近郊区作为另外两大层。考虑到各大层的层内方差与调查费用的差异,我们使用了一般情形的最优分配。而且各大层内的抽样也是考虑到各自的特点,采用了不同的抽样方法,不过基本上仍是分层(层内再分层)多阶抽样。在不属于区县管理的 12 个系统中,第一阶抽样采用了与中央单位抽样相类似的方法。总之,本案例是根据具体条件进行精心设计的一个范例。所以能做到这样,主要是得到主持单位的全力支持。如果没有这种支持,再理想的方案不能操作也是徒劳的。

5) 本案例的总体目标量估计及其方差估计也是严格与抽样方案配套的,其中包括了许多与 § 11.5 相类似的方法。值得一提的是在本案例中,在对总体两个总量 Y 与 X 的比值 R 的估计 \hat{R} 进行方差估计时,当应用 Taylor 展开将其化成关键的估计 $\text{Cov}(\hat{Y}, \hat{X})$ 时,采用了另一种简

单方法,即引进新指标 $u = x + y$, 将 u 按与 x 、 y 样本值完全相同的处理即可获得总和估计 \hat{U} 的方差估计 $v(\hat{U})$, 从而按 (3.12) 式即可获得 $\text{Cov}(\hat{P}, \hat{X})$ 的估计 $\text{Cov}(\hat{P}, \hat{X})$ 。

6) 本案例包括 833 个需要估计的总体目标量, 在按给定的数据公式处理后, 即可获得具体的估计量及其方差(实际上结果用标准差形式)数值。由此可对本项调查进行实际精度分析。在给定置信度 $1 - \alpha = 95\%$ 下, 可求得最大绝对误差 $d^* = u_{\alpha} s(\hat{\theta})$ 以及最大相对误差 $r^* = u_{\alpha} \text{cv}(\hat{\theta})$ 。本例中设计精度 $d = 2\%$, 若取 $u_{\alpha} \approx 2$, 则根据表 11.8, 全部目标量中有 97.3% 的实际 $d^* \leq d$ 。这个比例高于给定置信度 95%, 表明达到了事先要求的精度。这个结论也与设计时对 deff 的正确估计相吻合的。

§ 11.7 中国 1986 年 74 城镇人口迁移抽样调查^{*}

《中国 1986 年 74 城镇人口迁移抽样调查》(以下简称《迁移调查》)是由中国社会科学院人口研究所承担, 得到联合国人口活动基金资助, 被列为国家“七五”期间哲学和社会科学重点研究项目, 联合 16 省(市)人口研究单位共同合作研究《中国城镇人口迁移与城镇化》课题的组成部分。此项《迁移调查》填补了我国城镇人口迁移资料的空白, 提供了我国自 1949 年以来城镇人口迁移的流量、流向、结构、原因和后果的主要数据, 它们不仅是人口学、经济学、社会学、地理学、生态学等学科所需的基本数据资料, 也是国家决策部门制定改革政策的参考依据。

现已公布的计算机汇总数据资料(见本节参考资料 [1])是按城市规模汇总的实际样本数据。为进一步将这些宝贵的调查数据进行开发利用, 我们针对此项调查的抽样设计以及实际需要, 运用抽样调查的理论和方法提出了 74 城镇人口迁移有关目标量的估计方法以及对全国相应目标量的推总估计方法, 并用随机分组方法对 74 城镇上述目标量估计的精度(方差)进行了估计和分析, 同时还对全国指标的推算值做了评估。

一、抽样设计

《迁移调查》是在 1986 年 7 月开始进行的, 同年年底先后完成。其中 43 个城市的调查范围是居住在城市地区的人口, 即居住在城市市区、近

^{*} 本节正文节选自高嘉陵与冯士雍: 《中国 1986 年 74 城镇人口迁移抽样调查目标量估计方法与精度分析》, 原载《中国人口科学》1991 年第 3 期, 1~8。本文改正了原文中的若干印刷错误。

郊区、工业区的人口,它包括了城市中绝大部分非农业人口和一小部分农业人口;31个镇的调查范围是镇的总人口。以上城市的调查范围人口都已有明确的统计,我们以此做为《迁移调查》的目标总体。《迁移调查》确定的调查样本总量为25000户。各省(市)遵照大城市多抽,小城市和镇少抽的原则,以及根据本单位的工作条件和经费情况确定本省(市)城镇的样本量,从而决定了各城镇的抽样比(样本量与调查范围人口的比)。

《迁移调查》的抽样方案采用四级整群抽样。第一级抽样是从全国抽省(市),16个样本省(市)即16个人口研究单位所在省(市)是指定的,是根据研究单位的条件与可能自愿参加的。第二级抽样从上述样本省内抽城镇,并采用典型选取和随机抽取相结合的方法,在典型选取时按城市规模的大小,把城镇分为特大城市(100万以上人口)、大城市(50~100万人口)、中等城市(20~50万人口)、小城市(20万以下人口)和镇五类,选取中兼顾各种功能的城镇。第三级抽样是在城镇内抽取街道,抽取的方法是按比例分配分层,如城区、近郊区、工业区、商业区等层,对某些较小的城镇也有不分层情况。层内采用等概率或不等概率按地址编码系统抽样或简单随机抽样抽取街道。最后一级抽样是在被抽中的街道内用等概率系统抽样抽取家庭户(集体户划分为四人一群相当一户,集体户与家庭户的抽取比例按人口比例分配)。其中每一个街道抽取的户数也按该街道的总户数比例分配。对抽中的户则进行整户调查,即调查户内所有成员。

二、数据处理的基本思想和目标量的确定

以上《迁移调查》的抽样方案,从整体上说不是一个严格的概率抽样,特别是在省(市)一级和城、镇一级均未按概率抽样方法抽取,因而无法用抽样调查的一般方法处理,如根据样本对目标量做推总估计和精度估计。然而,我们注意到,这16省(市)已超过大陆当时29省(市、自治区)的半数,且东北、华北、华东、西北、中南、西南各地区内至少有2个省(市)。假若我们取消省(市)一级,直接观察74城、镇,从城镇的数量和地域分布来看对全国还有一定的代表性;并且各省(市)抽取的城镇是按同一原则典型选取的,因此若对调查的74城镇进行合理的“事后分层”,则可以利用分层抽样的计算方法对全国性指标进行数据处理。

鉴于以上理由,我们将全国居住在城市地区的人口做为推论总体,而将调查的74城镇中城市地区的人口作为目标总体。

据此我们分两步进行数据处理:第一步对每个调查的城镇计算有关

目标量的估计和相应的方差估计,然后汇总为目标总体相应的估计。第二步将推论总体和目标总体按同一原则分层,由74城镇目标总体主要目标量的估计分层加权得到全国城镇人口(推论总体)迁移指标的推算值。由于74城镇不是按概率抽样从全国抽取的,因此,对于推论值的偏差和精度则根据74城镇目标量估计与方差估计,做经验的定性分析。

在数据处理的第一步中,我们首先给出各城镇目标总体各类目标量的估计公式。对每个未知的目标量 θ ,用调查所得的样本数据对它进行估计,得到估计量 $\hat{\theta}$ 。由于 $\hat{\theta}$ 随样本而异,故有必要对它的精度加以讨论。描述一个估计量精度的准则之一是它的方差。方差表示估计量偏离其均值(对无偏估计量也就是目标量 θ 的真值)的大小的衡量,这种偏离在抽样调查中是不可避免的。如果我们用同一种抽样方法重复多次,即可得出方差的估计。当然在实际中重复抽样是不大可能的,因而根据样本数据作方差估计是十分重要的。在《迁移调查》中,由于采用的抽样方案比较复杂,且在各城镇中方法也不尽相同,因此在进行方差估计时,没有直接的公式可用,在本文中我们采用了随机分组法。

随机分组法亦称交叉子样本法,它的基本思路是将含有 n 个单元的样本(母样本)按一定方式划分为 b 个($b>2$)子样本(随机组),先分别求得每个子样本以及母样本目标量的估计,用不同子样本估计量之间的差异估计总体目标量的方差。随机分组方法的基本要求是这些子样本(随机组)的构成一般要求与母样本的抽样方法相一致,也就是说子样本的抽样结构与母样本的结构基本相同。

为了达到上述目的,我们将每个城镇中每个样本街道中的所有调查户按一定方法(详见下面第三小节)划分为 b 组。特大城市一般分为15组(上海分50组),大城市、中等城市、小城市分10组,镇分5组,以保证每一个街道小组中有5至10户的样本量。城镇中所有样本街道的第一组组成城镇的第一个子样本,所有的第二组组成第二个子样本,以此类推,这样将城镇母样本划分为 b 个子样本(随机组),分别对母样本及 b 个子样本进行数据处理,然后对每个城镇进行目标总体目标量的估计及其方差估计。

数据处理的第二步,首先将推论总体按目标总体的原则分层,然后计算推论总体目标量估计值。

我们将74城镇按地理分布(沿海、内地、边远地区)及城市规模(特大城市、大城市、中等城市、小城市及镇)共15层。

推论总体全国城镇的分层将在下面第四段《全国及各种规模城市和镇人口迁移指标的推算》中介绍。

《迁移调查》的指标项目共有 62 个之多, 从数据处理方法角度上讲, 这些指标的目标量可分为两类: 第一类是某个指标如 y 的总量 Y , 例如迁入人口总数, 迁入人口中男性总数等; 第二类是两个总量 Y 与 X 的比值 R , 例如迁入人口的性别构成, 即迁入人口男(女)性总数与迁入人口总数之比, 其中二者人口数都需要通过样本进行估计。其他如平均值或凡是在总人口 Z (调查时是已知的, 不需要估计) 中所占的比例 $P = Y/Z$ 则可归为第一类处理。我们参照《迁移调查》的研究报告(见参考资料[1]), 选择了以下指标为主要目标量, 它们包括了 $Y(P)$ 、 R 这两类目标量:

- (1) 城镇迁入人口占总人口的比例 P ;
- (2) 城镇迁入人口的性别构成 R_1 ;
- (3) 城镇迁入人口的年龄构成 R_2 ;
- (4) 城镇迁入人口的文化构成 R_3 ;
- (5) 城镇迁入人口的迁出地类型比重 R_4 ;
- (6) 城镇迁入人口的迁出年代比重 R_5 ;
- (7) 城镇迁入人口的迁入原因比重 R_6 ;
- (8) 城镇人口的分性别年龄构成 P_1 ;
- (9) 城镇人口的年龄构成 P_2 。

三、各城镇目标量的估计及其方差估计

《迁移调查》第三级抽样是在城镇中抽取街道, 有分层和不分层抽取两种情况。现按分层抽取介绍计算公式(不分层即层数为 1)。层内抽样又分二级, 第一级一般按等概率系统抽样抽街道, 在上海采用不等概率系统抽样。第二级按等概率系统抽样, 也即等距抽样, 从每个被抽中的街道中抽家庭户。对抽中的户则进行整户调查。由于街道和户的排列顺序是按地址编码和户籍顺序排列的, 与迁移情况无关, 也即与调查的指标量不相关, 故可看作是“随机排列”。在此情形, 系统抽样与简单随机抽样可看成是等价的, 因此我们可将这样的系统抽样按简单随机抽样公式处理。

1. 符号介绍

为介绍目标量的估计和方差估计方式, 首先引进若干记号:

h, i, j, k 分别为层、街道、户、人的编号; X, Y, \dots 为调查指标;

Z_{hi} 为 h 层第 i 街道第 j 户中的被调查人数;

m_{hi} 为 h 层第 i 街道抽中的户数;

n_h 为 h 层内抽中的街道数; M_i 为第 i 街道的总户数;
 N_h 为 h 层内街道总数; Z_{hi} 为 h 层第 i 街道的人数;
 Z_h 为 h 层内总人数; Z 为城镇人口总数.

2. 目标量的估计

首先我们讨论第 h 层某个指标 y 的总量 Y_h 的估计. 由于在每个街道内的抽样都是等概率的系统整群(户)抽样, 正如前面所指出的那样, 我们可用简单随机抽样的有关公式. 对于 h 层内第 i 街道的指标 y 的平均数 \bar{Y}_{hi} 可用以下简单估计:

$$\hat{\bar{Y}}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} \sum_{k=1}^{Z_{hi}} y_{hijk}, \quad (11.67)$$

其中 y_{hijk} 是第 h 层第 i 街道第 j 户第 k 人的指标.

因此 h 层内总量 Y_h 可按以下公式估计:

$$\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi} \hat{\bar{Y}}_{hi}. \quad (11.68)$$

若层内抽样是按街道人口数成比例的不等概率系统抽样, 则按照不放回不等概率抽样的 Horvitz Thompson 估计:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \frac{M_{hi} \hat{\bar{Y}}_{hi}}{\pi_i} = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{Z_{hi}}{\pi_{hi}} M_{hi} \hat{\bar{Y}}_{hi}, \quad (11.69)$$

其中 $\pi_i = n_h \frac{Z_{hi}}{Z_h}$ 是第 i 街道被抽中的概率.

获得层内总量 Y_h 的估计后, 比例 $P_h = Y_h/Z_h$ 的估计即可随之得到:

$$\hat{P}_h = \frac{\hat{Y}_h}{Z_h}, \quad (11.70)$$

而两个量的比 $R_h = Y_h/X_h$ 的估计则由下式给出:

$$\hat{R}_h = \frac{\hat{Y}_h}{\hat{X}_h}, \quad (11.71)$$

其中 \hat{X}_h 是指标 x 的层总量的估计, 可以从 x_{hijk} 按公式(11.68)或(11.69)同样处理.

当在层内抽中街道的总户数以及在不同街道抽中的户数都是按户数比例分配时, 则样本是自加权的, 此时目标量的估计可简化为:

$$\hat{Y}_h = \frac{1}{f_h} \sum_i \sum_j \sum_k y_{hijk} \quad (11.72)$$

其中 f_h 为层内总抽样比.

因而

$$\hat{P}_h = \frac{1}{f_h} \sum_i \sum_j \sum_k y_{hijk} / Z_h, \quad (11.73)$$

$$\hat{R}_h = \sum_i \sum_j \sum_k y_{hijk} / \sum_i \sum_j \sum_k x_{hijk}. \quad (11.74)$$

即目标总量的估计等于样本指标总和除以抽样比; 比例型估计为样本总和与层内总人数之比; 比值型估计为两个指标的样本总和之比.

根据分层抽样公式, 城镇目标总量的估计为:

$$\hat{P} = \sum_{h=1}^L \hat{P}_h, \quad (11.75)$$

其中 L 是城镇中划分的层数.

比例型目标量 P 及比值型目标量 R 分别可估计为:

$$\hat{P} = \sum_{h=1}^L W_h \hat{P}_h, \quad (11.76)$$

$$\hat{R} = \sum_{h=1}^L W_h \hat{R}_h, \quad (11.77)$$

其中 $W_h = Z_h/Z$ 是层权.

3. 估计量方差的估计

正如在第二段中所述, 我们采用随机分组法对估计量的精度(用方差表示)进行估计. 我们首先介绍随机组的组成方法, 然后根据随机组给出方差的估计. 若在城镇中不分层抽取街道, 则将每个被抽中的街道中的(设为 m 个)家庭户用系统抽样方法划分为 b 个随机组. 在整数 1 至 b 中, 抽取一个随机整数 r , 将第一个样本户划为第 r 组, 第二个样本户为第 $r+1$ 组, 以此类推, 直到某一样本户为第 b 组, 以下的样本户顺序为第 1 组, 第 2 组, \dots , 第 r 组, \dots , 第 b 组, 再从第 1 组顺序排下去. 如果街道的样本量 m 不是 b 的整倍数, 令 $m = bO + q$ (O 为整数), 则余下的 q ($q < b$) 个样本分别划为 r_1, r_2, \dots, r_q 组, r_1, \dots, r_q 为从 1 至 b 个整数中抽取的 q 个不放回的随机整数. 第 α 个随机组则由所有 n 个样本街道的第 α 组的家庭户组成.

对于第 α 个随机组, 采用上述目标量估计公式计算某目标量 θ (Y, P 或 R) 的估计值 $\hat{\theta}_\alpha$, 另外采用未分组的母样本按照上述公式求得的 θ 的估计量为 $\hat{\theta}$, 则 $\hat{\theta}$ 的方差的随机分组估计量为:

$$v(\hat{\theta}) = \frac{1}{b(b-1)} \sum_{\alpha=1}^b (\hat{\theta}_\alpha - \hat{\theta})^2. \quad (11.78)$$

若城镇采用分层抽取街道, 则将 L 个层按以上方法每层分为 b 组, 对城镇和每个随机组都进行总量估计, 有:

$$\hat{P} = \sum_{h=1}^L \hat{P}_h, \quad (11.79)$$

$$\hat{P}_\alpha = \sum_{h=1}^L \hat{P}_{h\alpha}. \quad (11.80)$$

则目标量 Y 估计的方差估计为:

$$v(\hat{P}) = \frac{1}{b(b-1)} \sum_{\alpha=1}^b (\hat{P}_\alpha - \hat{P})^2. \quad (11.81)$$

对比例型估计 P , 我们用下式估计 P_α 及 P :

$$\hat{P}_\alpha = \hat{P}_\alpha / Z, \quad \hat{P} = \hat{P} / Z, \quad (11.82)$$

于是 $V(\hat{P})$ 可用下式估计:

$$v(\hat{P}) = \frac{1}{b(b-1)} \sum_{\alpha=1}^b (\hat{P}_\alpha - \hat{P})^2. \quad (11.83)$$

比值型估计 \hat{R} , 我们仍用同样的估计量:

$$\hat{R} = \hat{P} / \hat{X}, \quad (11.84)$$

这里 \hat{P} 与 \hat{X} 根据(11.79)式计算.

为了估计 $V(\hat{R})$, 我们利用泰勒级数, 可以得到 $V(\hat{R})$ 的以下近似公式:

$$V(\hat{R}) = \hat{R}^2 \left[\frac{V(\hat{P})}{\hat{P}^2} + \frac{V(\hat{X})}{\hat{X}^2} - \frac{2\text{Cov}(\hat{P}, \hat{X})}{\hat{P}\hat{X}} \right],$$

它的一个估计是:

$$v(\hat{R}) = \hat{R}^2 \left[\frac{v(\hat{P})}{\hat{P}^2} + \frac{v(\hat{X})}{\hat{X}^2} - \frac{v(\hat{U})}{\hat{P}\hat{X}} \right], \quad (11.85)$$

其中 $v(\hat{P})$ 及 $v(\hat{X})$ 用(11.81)式计算, 而 $v(\hat{U})$ 则对新指标 U , 即 $u_{hjk} = y_{hjk} + \alpha_{hjk}$ 用公式(11.81)式计算而得.

在求得 34 城镇目标量估计和方差估计之后, 我们按照前述的分层方法, 用分层抽样公式得到目标总体 74 城镇主要目标量的估计和方差估计.

四、全国及各种规模城市和镇人口迁移指标的推算

全国及各种规模城镇目标量的估计是通过计算了 74 城镇的目标量估计之后, 对它们进行“事后分层”用分层抽样公式求得的. 为此, 我们需对全国城镇分层, 并需已知各层中的城市 and 居住在这些城市地区的层人口数. 将全国城镇分层的原则必须与 74 城镇分层原则一致. 因为居住在城市地区的人口包括了城市中绝大部分的非农业人口, 所以在对全国城市按规模分层时, 我们使用公安部所编 1986 年度全国分县市人口统计资料(见参考资料[4])中市非农业人口一览表. 在确定某规模层的城市时, 从大到小取到在本规模层中被调查城市是非农业人口最少的城市为止.

比如全国的大城市,我们从福州市取到株洲市,全国的中等城市从双鸭山市取到肇庆市,也因为居住在城市地区的人口包括了城市中绝大多数的非农业人口,所以我们以43城市的调查范围人口与城市中非农业人口的比例为权数,来计算各层居住在城市地区的人口数。经分层加权计算得推论总体全国居住在城市地区的总人口为14.963千万人。镇的调查范围与它的总人口一致为20.37千万人。

据统计,1986年全国城市总人口为2亿3千万人,非农业人口为1亿2千万人,由《迁移调查》推算的全国居住在城市地区的人口为1亿5千万人。那么居住在全国城市地区的农业与非农业人口的比例约为1:4,我国城市中农业与非农业人口比例的变化与城市建制原则和城乡划分标准的变动有关。《迁移调查》的调查范围是在城市地区的实际人口,在一定时期内其农业与非农业人口的比例是相对稳定的,因此由《迁移调查》所推算的全国城市地区一亿五千万人的人口数在一定程度上反映了中国城市化的实际水平(考虑到一些特大城市没有调查郊区,城市化真实水平农业人口的比例要略高些)。

全国各层目标量的估计由层内抽中城镇的目标量估计值按居住在城镇地区人口数加权求得:

$$\hat{\theta} = \sum_{d=1}^c W_d \hat{\theta}_d = \sum_{d=1}^c \frac{Z_d}{Z} \hat{\theta}_d, \quad (11.86)$$

其中: C 为层内抽中城市数;

Z_d 为层内抽中第 d 个城市居住在城市地区的人口;

Z 为层内居住在城市地区的总人口。'

全国各地区及各种规模城镇主要人口迁移指标的推算由各层目标量估计按分层抽样公式求得(以下略)。

参 考 资 料

- [1] 《中国1986年74城镇人口迁移抽样调查资料》,《中国人口科学》专刊 II, 1988年。
- [2] Kish L. Survey Sampling, John Wiley Sons, 1965.
- [3] Wolter K M. Introduction to Variance Estimation, Springer Verlag 1985.
- [4] 《中华人民共和国全国分县市人口统计资料》1986年度中华人民共和国公安部编, 地图出版社, 1987年。

评 注

- 1) 本案例的调查事先未经过严格的抽样设计, 前两级(阶)抽样是非

随机的,甚至是人为的,但“挑选”出来进行调查的74个城镇在地理上分布还是比较均匀的,从而对全国仍有一定的代表性。第三、四级抽样则基本上还是严格的,尽管对每个城镇并不采用同样的方法,严格地说,本调查只对所调查的74个城镇即文中所说的“目标总体”有意义,而对全国城镇(人口)即文中的“推论总体”只有参考意义。

2) 本案例设计时未考虑到对总体目标量的估计,更谈不上精度估计。参考资料[1]中列出并进行讨论的只是样本汇总资料。这种情况在前几年全国性大型调查中并不少见,在今天许多报刊上发表的市场调查或公众调查更是常见的。如果样本是自加权的(遗憾的是在这类调查中这极为少见),样本比例或平均数才可作为总体相应目标量的(无偏)估计。否则,样本值必定与总体目标量之间存在或多或少的偏倚。因此,根据样本资料必须进行推总估计,这应该引起每项抽样调查主持或决策者所重视的。

3) 作为补救措施,本例用随机分组法将样本分成若干随机组(子样本)。分组方法是使子样本的结构与原母样本一致。用一种确定的方法来估计每个子样本的(对总体目标量)估计,再求其平均值作为正式估计,最后用(11.78)式来计算该估计量的方差。当然随机组法的效率不够高,所得估计量的方差估计不够精确。但优点是计算并不复杂。若需进一步提高效率,还可使用第9章中介绍的其他方法,例如平衡半样本方法、Jackknife方法或Bootstrap方法来估计方差,当然,这需要以付出更大的计算量为代价。

§ 11.8 中国妇女社会地位调查^{*}

一、调查方案

1. 调查目的

中国妇女社会地位调查的目的有四:第一,客观、准确、系统地描述中国妇女社会地位的现状与发展;第二,分析研究妇女社会地位变迁的规律和影响因素;第三,进行省际及不同层次妇女地位比较,并在可能的情况下进行国际比较;第四,总结、筛选评价妇女社会地位的综合评价指标,以

^{*} 本项目由中华全国妇女联合会与国家统计局联合主持,本节正文摘选自陶春芳、蒋永萍主编的《中国妇女社会地位概观》第二章研究方法,中国妇女出版社,1993。该章原文由蒋永萍、胡忠兵、杨李执笔,其中抽样是由胡忠兵设计的。

便进行长期监测。

围绕研究目的,本项目的研究设计遵循以下原则:

(1) 以男性为参照系,同男性比较,进行跨阶层、跨地域、跨职业的考察。妇女地位是相对于男性而言的,没有男性地位也就无所谓女性地位的探讨。

(2) 考察妇女的总体地位,而不是单个人的特殊状态。即妇女地位的研究是就整体而言的。

(3) 以当代妇女地位为主,同时为说明发展,还要以女性自身的过去为参照系,做不同年代妇女地位的比较。

(4) 考察中国妇女社会地位,为此在研究方法、指标设计上要体现中国特色,同时为取得国际社会的同一认识及相互比较的需要,也注意借鉴国外妇女研究方法和指标。

2. 调查内容、指标

中国妇女社会地位调查的指标体系依据我国社会经济发展现状,参照联合国及亚太地区监测妇女地位的指标设置,它包括以下八方面内容:

(1) 法律权利; (2) 生育与健康; (3) 教育; (4) 劳动就业; (5) 社会参与与政治参与; (6) 婚姻家庭; (7) 自我认知与社会认同; (8) 生活方式。各项内容的主要指标从略。

3. 调查方法及调查表

中国妇女社会地位调查主要采取三种方法: 个人问卷调查; 社区及企事业单位直接统计调查; 统计文献调查。

(1) 个人问卷调查: 即“中国妇女社会地位调查个人问卷”, 此调查表采用调查员入户访谈方法, 内容包括中国妇女社会地位调查的各个方面, 目的是从 18~64 岁男女公民的亲身经历、行为、观念、体会中了解中国妇女社会地位各个层面的历史与现状。个人调查问卷是中国妇女社会地位调查的主调查表, 调查标准时点为 1990 年 9 月 15 日。

(2) 社区及企事业单位直接统计调查: 所谓直接统计调查, 是指调查员深入到被调查单位, 使用调查表向有关部门搜集数据、了解情况的方法, 其特点是把一个组织或社区作为研究对象, 从而把握整体的结构、性质, 以作为个人问卷调查的背景资料、补充资料和校验资料。本次调查采用直接统计调查的调查表有 5 种:

① 村民委员会调查表, 重点了解农村整群妇女人口、教育、劳动等情况。

② 工业企业调查表, 重点了解城镇工业企业整群妇女的职业分布、劳动保护、劳动报酬、劳动效率等情况。

③ 高等院校调查表, 重点了解分学科男女毕业生情况。

④ 产院、福利院调查表, 重点了解被遗弃女儿童少年的情况。

⑤ 县以上机关团体调查表, 重点了解妇女的社会参与、法律权利等情况。

(3) 统计文献调查: 所谓统计文献调查, 是指利用有关部门的现成统计文献而进行的汇总统计调查。本次调查使用的统计文献调查表依据国家统计局以及劳动人事、教育、司法、卫生、计划生育等部门的现有统计资料编制。其目的是从现有的统计资料中获取与妇女地位有关的各种信息, 把握影响妇女社会地位变化的宏观背景和中国妇女社会地位发展的历史脉络与总体概况。

4. 调查的组织实施

中国妇女社会地位调查是在全国妇联和国家统计局的领导下, 由全国妇联妇女研究所和国家统计局社会司具体组织实施的。调查的研究设计等前期准备、全国性调查的组织实施、数据汇总和国家级报告的撰写, 由全国妇联妇女研究所《中国妇女社会地位调查》课题组在国家统计局社会司有关同志协助下完成的。各样本省、直辖市、自治区子课题组负责本省、直辖市、自治区调查工作的组织实施和地区性报告的撰写。

调查员: 本调查全部调查员由妇联系统干部担任, 23个省、直辖市、自治区共投入调查员 2000 余人。为确保调查质量, 疏通工作环节, 在各样本县、市聘任了调查指导员。调查员的培训分两级进行。《中国妇女社会地位调查》课题组负责培训各省、直辖市、自治区的调查研究人员。各省、直辖市、自治区的全部调查员和调查指导员由省、直辖市、自治区子课题组负责。

试点: 为保证调查用表的可操作性, 正式调查前, 组织了两次试点。第一次为 1989 年 9 月在内蒙古包头市, 第二次为 1990 年 8 月在北京市怀柔县。试调查中发现的问题经专家及本课题研究人员反复讨论修改后定稿。

质量检验: 质量是调查的生命。为确保质量, 减少调查误差, 本调查建立了严格的质量检验制度。问卷的质量检验的内容包括回收率、可信度、有效性、抽样四个方面, 并分四级进行: ① 调查员自检。② 市县调查指导员对报送的各式调查员进行百分之百的复核与检验, 发现差错和漏

填,退回调查员回访查实。③各省、直辖市、自治区抽验全部调查问卷、表的20%(实施中很多省检验比例达100%)。④全国妇联妇女研究所课题组于1991年1月对北京等11个省、直辖市回收的各式调查表进行了2%的抽验。

数据录入、整理:中国妇女社会地位调查个人调查问卷的数据处理采取统一标准、分级录入,两级汇总分析的方法,个人问卷原始数据的计算机录入由各省按全国统一编制的录入程序分别完成,数据的汇总分析由全国和各省使用SPSS统计分析软件同时进行。其他各式调查表由全国和各省分头录入并分析。数据录入后,本课题研究人员除按逻辑关系对全部数据进行机械清理外,还抽验、清理了部分数据的所有记录,以消除由于录入等前期工作造成的差错。

二、抽样方法及抽样结果评估

1. 抽样方法

本次调查个人问卷样本所推断的总体定义为:调查标准时点上所有参与调查的省、自治区和直辖市居住在家户内18岁及以上,64岁及以下全体男女公民。抽选样本时,以家庭户做为基本抽样单位,在每个样本户内18周岁至64周岁的两性人口中,按特定随机程序抽选出一人做为调查对象。

为了使样本不但对整个研究总体进行推断,同时也能分别对每个参与调查的省级子总体进行独立推断,并进行省际比较,本次调查以省做为研究域,每省抽取相等规模的样本。

抽样组织形式为,在各省内使用统一的分域、分层、多阶段、概率比例(PPS)、随机等距的抽样方式。

由于我国城乡差异大,且农业人口所占的比重高,对妇女地位而言,农村的同质性大于城市,从调查费用及难度来看,农村又明显高于城市。为了能分析比较城乡差别,提高抽样精度,并能保证城市分析具有足够的样本量,省内进一步按城乡分域(实际上是作为研究域的层)。这里的城乡分域是按非农业或农业户口划分的。城市域主要指非农业人口,包括城市非农业人口以及县镇非农业人口;农村域则不仅包括县属农村的农业人口,还包括城市所属和城市辖区的农业人口。各省城乡两域的样本规模相等,据此做省级测算时,要求对数据结果进行加权处理。

样本量采用保守的方法确定。要求在各省内城市域或农村域中能分别以95%的可信度保证百分比的绝对误差不超过5%,因此省内分域的

简单随机抽样的平均样本量为:

$$n_0 = \frac{t^2 p(1-p)}{d^2}.$$

取 $p=0.5$, $t=1.96$, $d=0.05$, 则可得到对 n_0 的保守估计: $n_0=385$.

求得简单随机抽样的样本容量后, 用设计效应进行调整, 从而得出复杂抽样设计所需的样本量. 按照经验, 类似抽样的设计效应 $\text{deff} \leq 2$, 为保险起见, 设 $\text{deff}=2.5$, 从而省内分域的样本量为:

$$n' = \text{deff} \times n_0 = 2.5 \times 385 = 963.$$

故取 $n_1=1000$ 为省内分域的样本量.

鉴于以上各步推导都是保守的, 因而由调查结果所计算的精确度会更高. 随着样本量的增加, 作省级分析与作全国分析时精确度还会高.

为了降低抽样误差, 提高抽样精度, 在各省内还按城乡分域对初级抽样单位进行分层. 分层按地理、经济及人口规模等进行, 以使同一层内的初级抽样单位具有尽可能高的同质性. 其中城市域按规模大小及历史状况分为三层: 大型城市(50 万人口以上)、中小型城市(50 万人口以下的非新建市, 即 1987 年以前建制的城市)、新建县级市和县. 农村按地理条件分为三层: 丘陵县(或城市辖区)、山区县(或城市辖区)、平原县(或城市辖区). 这里的县包括新建县级市. 为避免过多的加权计算, 使用与规模大小成比例的概率抽选各级抽样单位. 域内各层中样本量的分配也与层的规模大小成比例, 即在各层中使用相同的抽样比: $f_{\text{城层}}=f_{\text{城域}}$, $f_{\text{乡层}}=f_{\text{乡域}}$.

对参与调查的各省, 均采用四个阶段抽样:

第一阶段: 各省内分城、乡域均以县市作为初级抽样单位(对自代表单位, 将其挑出, 单独为层, 并按相同的抽样比分配应抽样本数). 经权衡, 在各个域共抽 25 个初级抽样单位, 全省共 50 个初级抽样单位, 其一半属城市域, 一半属农村域. 抽选方法是先将省内分域对所有初级抽样单位分层, 然后在同一域内以相同的抽样比计算出各层应抽样本户数, 再以此数除以 40(每个初级抽样单位内平均应抽户数), 得出各层内应抽的初级抽样单位数. 同一域内各层的初级抽样单位数之和应为 25, 全省的初级抽样单位数之和为 50. 最后在各层内, 以概率比例方法抽出所需数目的初级抽样单位.

第二阶段: 在每个样本市、县内, 用概率比例方法抽选经计算后确定数目的街道(或乡). 一般一个初级抽样单位抽 1 个街道(或 2 个乡), 个

别规模大的初级抽样单位可能不止抽这个数(这里特指将自代表单位单独挑出作层并以相同的抽样比分配样本数目的情况)。由于初级抽样单位按城乡分域交叉,所以有些市、县的可能街道或乡都要抽,有些则只抽其一。

第三阶段:在每个街道(或乡)内,用概率比例方法抽选出2个居(或村)民委员会。

第四阶段:在每个样本居(或村)委会内,根据第四阶段抽样比和调查时点上的实际户数,计算出应抽户数。将居(或村)委会内所有家庭户列表,按简单随机(或等距)抽样方法抽出样本户。每个居(或村)委会平均应抽20(或10户)。这样做是为了提高抽样精度,考虑到城市域中居委会内样本的异质性大于农村域中村委会内样本的异质性,而且居委会的规模有很多都大于村委会的规模等因素。

这样,每个域的计划样本量为四个阶段抽样单位数的乘积:

农村域: $25 \times 2 \times 2 \times 10 = 1000$ (户),

城市域: $25 \times 1 \times 2 \times 20 = 1000$ (户),

全省总的样本量为城乡两域之和:

$$1000 + 1000 = 2000 \text{ (户)}.$$

域内总的抽样比为各阶段抽样比的乘积:

$$f = \frac{a \text{Mos}_a}{\sum \text{Mos}_a} \times \frac{b \text{Mos}_{a\beta}}{\text{Mos}_a} \times \frac{c \text{Mos}_{a\beta\gamma}}{\text{Mos}_{a\beta}} \times \frac{d^*}{\text{Mos}_{a\beta\gamma}},$$

其中: $a = 25$ 为第一阶段抽样单位数;

$b = 1$ (或2) 为第一阶段样本单位内平均抽出的第二阶段样本单位数;

$c = 2$ 为第二阶段样本单位内平均抽出的第三阶段样本单位数;

d^* 为每个居(或村)委会内计划抽选的户数, d^* (城市域) 20, d^* (农村域) = 10;

Mos_a 为分域抽样框上市、县户数;

$\text{Mos}_{a\beta}$ 为分域抽样框上街道(或乡)户数;

$\text{Mos}_{a\beta\gamma}$ 为分域抽样框上居(或村)委会户数。

使用概率比例方法时,如果最后阶段抽样框上的居(或村)委会户数与调查时点上的实际户数不符,则应抽户数为:

$$d = N_{a\beta\gamma} \times f_4 = N_{a\beta\gamma} \times \frac{d^*}{\text{Mos}_{a\beta\gamma}},$$

其中: $f_4 = d^*/MOS_{asy}$ 为第四阶段抽样比;

N_{asy} 为调查标准时点上居(或村)委会的实际户数.

若最后阶段使用随机起点的等距抽样, 只要使用抽样间隔: $k = 1/f_4$, 进行等距抽样, 则实际抽出的户数就符合上式要求.

为了使从家庭户中抽出的被调查人所组成的样本在年龄、性别等方面的分布与总体分布尽可能一致, 采用下述特点的抽样方式进行户内抽人(见 leslie kish: «Survey Sampling»). 这一抽样程序的关键在于写出序号, 并作选择. 序号的排法是男性在前、女性在后. 男性中以最年长的排在第一位, 次年长的排在第二位, 以此类推, 女性的最长者排在男性的最幼者后面, 余下的排列与男性相同. 按这种排列顺序作序号, 如 1, 2, 3, 4 等, 填入序号栏内. 调查员根据手持的“×式选择表”, 按照家庭户的人数多少作选择.

选择表的格式共有 A 、 B_1 、 B_2 、 C 、 D 、 E_1 、 E_2 、 F 八种, 分别占总调查表中 $1/6$ 、 $1/12$ 、 $1/12$ 、 $1/6$ 、 $1/6$ 、 $1/12$ 、 $1/12$ 、 $1/6$, 其具体形式分别如表 11.10 所示.

表 11.10

A 式选择表		B ₁ 式选择表	
如果家庭户中 18 岁至 64 岁人口数为	被抽选人的序号为	如果家庭户中 18 岁至 64 岁人口数为	被抽选人的序号为
1	1	1	1
2	1	2	1
3	1	3	1
4	1	4	1
5	1	5	2
6 或以上	1	6 或以上	2

B ₂ 式选择表		C 式选择表	
如果家庭户中 18 岁至 64 岁人口数为	被抽选人的序号为	如果家庭户中 18 岁至 64 岁人口数为	被抽选人的序号为
1	1	1	1
2	1	2	1
3	1	3	2
4	2	4	2
5	2	5	3
6 或以上	2	6 或以上	3

D 式选择表	
如果家庭户中 18 岁至 64 岁人口数为	被抽选人的序号为
1	1
2	2
3	2
4	3
5	4
6 或以上	4

E ₁ 式选择表	
如果家庭户中 18 岁至 64 岁人口数为	被抽选人的序号为
1	1
■	2
3	3
■	3
5	3
6 或以上	5

E ₂ 式选择表	
如果家庭户中 18 岁至 64 岁人口数为	被抽选人的序号为
1	1
2	2
3	2
4	4
5	5
6 或以上	5

F 式选择表	
如果家庭户中 18 岁至 64 岁人口数为	被抽选人的序号为
1	1
2	2
3	3
4	4
5	5
6 或以上	6

抽选时, 如果户内 18 至 64 周岁的人口数大于 6, 则按选择表, 若中选人的序号为 1, 则加选序号为 7 的家庭成员; 若中选人的序号为 2, 则加选序号为 8 的家庭成员, 以此类推。当加选的序号大于户内被排序的人口数时, 只选择表中给定的一人调查即可。

按这种户内抽人法抽选被调查者, 当样本量足够大时, 样本在年龄及性别方面的分布将与总体分布一致。

根据前述的抽样方式进行抽样, 总体均值的估计形式为:

省内分域: 由于 PPS 抽样估计总体均值具有自行加权的特点, 因而省内分域的总体均值的无偏估计量就是样本均值。

各省均值的估计: 因为各省内将城、乡分别作为是研究域的层, 且各层内的抽样比不同, 但所抽取的样本量相同, 因而各省均值的无偏估计为:

$$\bar{y}_s = W_s \bar{y}_s + W_{ns} \bar{y}_{ns}$$

其中, W_s 与 W_{ns} 分别为省内非农业人口与农业人口所占的比例, \bar{y}_s 与 \bar{y}_{ns} 为分域的样本均值。

全国均值的估计: 首先, 依据社会经济地理特征将全国 29 个省、自

治区、直辖市(海南仍归为广东)分为沿海省市、内地省份、边远省区三大层,又依据人均国民收入与平均受教育年限两项指标综合加权平均为社会经济发展相对指数,做出层内分层标志,将每层分为两个小层。从中抽选出11个省、直辖市做为推断全国的样本省。这11个省、市为:北京、江苏、河北、广东、吉林、湖北、安徽、江西、青海、甘肃、贵州。由这些省的数据推断全国时亦应作适当加权处理:

$$y_{\text{全国}} = \sum_{i=1}^{11} W_i y_i.$$

这里的 W_i 由层规模大小及样本省规模大小共同决定(具体数值见第三段)。这一估计量也是无偏的。

2. 抽样误差与统计推断

抽样调查的目的是通过样本统计量来估计总体参数,误差是指样本统计量与未知总体参数之间的差异。产生误差的原因主要有三种:即登记误差、系统误差与随机误差。从理论上讲,前两种误差都是可以克服的,但只要是用样本来代表总体,随机误差就一定存在。因为随机误差是由于抽样时的各种随机因素造成的,是用部分来代表全体所必然产生的误差。随机误差又分为抽样实际误差和抽样平均误差。抽样实际误差是指实际抽出一个样本后,样本统计量与对应总体参数之间的随机误差。抽样平均误差是所有可能出现的样本统计量与对应的总体参数之间的平均误差程度,从同一总体中抽取的单位数相同的所有样本都具有同一个抽样平均误差,简称抽样误差,也就是统计量抽样分布的标准差。在无偏的情况下,抽样误差反映的是样本统计量的波动程度,即抽样的精确程度,因而抽样误差越小,抽样的精度也就越高。

就本次调查而言,考察样本对总体的代表性也从抽样实际误差及抽样误差这两方面进行。本次调查的基本抽样单位是家庭户,被调查者是从家庭户中按特定的随机程序抽选的,样本的年龄、性别分布只与户内抽人的程序有关,与抽选样本家庭户的抽样方法无关。除年龄、性别外,其它变量的抽样误差都直接取决于样本家庭户的抽选办法。

对总体参数已知的部分变量,我们直接将全国数据样本值与总体值(1990年第四次人口普查数据)进行比较。各项抽样实际误差如表11.11所示。

从表11.11可以看出:在有关总体基本情况的变量上,抽样实际误差有大有小,但总的来说,样本分布的趋势与总体趋势一致。

表 11.11

性 别	样 本 比 例	总 体 比 例	抽样实际误差
男	52.1%	51.78%	0.3%
女	47.9%	48.22%	0.3%
年 龄	样 本 比 例	总 体 比 例	抽样实际误差
18~19 岁	4.8%	6.5%	1.7%
20~24 岁	12.6%	17.6%	5.0%
25~29 岁	16.6%	17.4%	0.8%
30~34 岁	13.8%	10.5%	3.3%
35~39 岁	16.9%	12.9%	4.0%
40~44 岁	11.9%	9.8%	2.1%
45~49 岁	7.2%	7.3%	0.1%
50~54 岁	6.4%	6.5%	0.1%
55~59 岁	5.4%	6.2%	0.8%
60~64 岁	4.4%	5.3%	0.9%
文 化 程 度	样 本 比 例	总 体 比 例	抽样实际误差
不识字或识字很少	20.0%	20.6%	0.6%
小学	33.5%	42.3%	8.8%
初中	29.8%	26.5%	3.3%
高中	11.8%	7.8%	4.5%
中专	2.6%	1.7%	0.9%
大专	1.5%	1.0%	0.5%
大学本科及以上	0.8%	0.6%	0.2%

对总体参数未知的部分变量,需计算其抽样误差。鉴于本次调查用到多种抽样方法,因此抽样误差的计算应将每种方法所对应的误差估计公式结合起来使用:

(1) 在求比例的估计时,用二项式公式来估计 $V(p)$,则由样本得到的 p 的方差的无偏估计为:

$$V(p) = \frac{N-n}{N(n-1)} p(1-p).$$

而利用整群抽样技术来计算 $V(P)$ 的值, 则公式为:

$$V(p) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^n (P_i - p)^2.$$

这两个公式在多阶段抽样中的不同阶段使用, 对于分层的情况, 将在各层中独立使用.

(2) 在求数值的估计时, 由样本得到的 \bar{y} 的方差的无偏估计为:

$$V(\bar{y}) = \frac{N-n}{Nn(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2.$$

(3) 利用分层误差计算各阶段误差, 对含有分层的某一特定阶段的抽样误差公式为:

$$V(p) = \sum_{k=1}^L W_k^2 V(p_k),$$

其中: L 为层数, W_k 为层的权, p_k 为层的样本比例, $V(p_k)$ 是指在第 k 层反复取样时 p_k 的方差. 这一公式对求数值的公式也适合, 只须将 p 改为 \bar{y} , p_k 改为 \bar{y}_k 即可.

(4) 计算多阶段抽样的误差: 本次抽样全国样本的抽选是五阶段的, 其抽样误差计算公式为:

$$\begin{aligned} V(p) = & V_1\{E_2[E_3(E_4(E_5(p)))]\} + E_1\{V_2[E_3(E_4(E_5(p)))]\} \\ & + E_1\{E_2[V_3(E_4(E_5(p)))]\} + E_1\{E_2[E_3(V_4(E_5(p)))]\} \\ & + E_1\{E_2[E_3(E_4(V_5(p)))]\}. \end{aligned}$$

这里 V 表示方差, E 表示均值, 下标表示各具体阶段数. 由于多阶段抽样的误差几乎集中在前一、两个阶段, 因而实际计算抽样误差时将最后阶段的误差忽略.

下面即是此次调查部分变量的抽样误差及估计均值:

问 1: 您一共上了几年学?

样本均值: 5.7 年, 抽样误差: 0.36 年.

问 32: 您的初婚年龄? (调查表中共有常用汉字 125 个)

样本均值: 22.41 岁, 抽样误差: 0.34 岁.

问 68: 您的识字个数? (调查表中共有常用汉字 125 个)

样本均值: 82.30 个, 抽样误差: 4.76 个.

问 58: 在您周围是否存在如下男女不平等的现象(表 11.12)? (按感受的强烈程度选两项.)

表 11.12

选 择 项	首选比例	抽样误差	次选比例	抽样误差
1. 招生时男女分数线不平等	9.4%	1.54%	1.6%	1.59%
2. 男女就业机会不均等	14.5%	1.86%	7.1%	1.34%
3. 男女同工不同酬	7.2%	1.54%	5.3%	1.21%
4. 妇女被裁减下来的多	1.6%	1.34%	3.1%	0.94%
5. 女性离婚再婚难	7.8%	1.76%	4.3%	1.33%
6. 妇女受侮辱、诽谤多	9.6%	1.74%	7.1%	1.59%
7. 生女孩的女性被歧视	22.1%	2.45%	21.5%	2.49%
8. 女儿继承遗产难	6.1%	1.43%	21.7%	2.52%
9. 都不存在	20.5%	2.45%	1.8%	1.40%

三、数据处理方法

中国妇女社会地位调查个人调查问卷全国性数据的处理利用目前国际上常用的统计软件 SPSS 完成。目前主要进行了单变量分析和双变量交互分析两类统计。

数据计算分总体、男性总体、女性总体、城镇总体、农村总体、城镇男性、城镇女性、农村男性、农村女性九个域进行。为了使结果具有更科学、更全面、更概括的说明价值,统计时对上述九个域分别做了加权处理。

对城镇总体、农村总体、城镇男性、城镇女性、农村男性、农村女性六个域使用了抽样时产生的省级权数,见表 11.13。

表 11.13

省(市)名	权 数 值	省(市)名	权 数 值
北 京	0.017175323	安 徽	0.205197483
江 苏	0.107713136	江 西	0.139544496
河 北	0.116748468	青 海	0.040260095
广 东	0.1348751599	甘 肃	0.03444177
吉 林	0.048751599	贵 州	0.049161759
湖 北	0.106149611		

对男性总体、女性总体。由于抽样设计时,城镇样本量与农村样本量相等,但实际上各省(市)农村人口与城镇人口的比例并不相等,而且城乡两域问卷的回收量也不均等,因此给定的权数必须考虑对城乡样本的各

自修正。假设某省回收的问卷总量为 S , 城镇数为 $S_{\text{城}}$, 农村数为 $S_{\text{乡}}$, 省级权数为 $W_{\text{省}}$, 根据 1990 年第四次人口普查 10% 抽样汇总资料, 得到某省的城镇人口百分比为 $H_{\text{城}}$, 农村人口百分比为 $H_{\text{乡}}$, 该省的城镇权数与农村权数分别为:

$$W_{\text{城}} = \frac{S \times H_{\text{城}}}{S_{\text{城}}} \times W_{\text{省}}, \quad W_{\text{乡}} = \frac{S \times H_{\text{乡}}}{S_{\text{乡}}} \times W_{\text{省}}.$$

表 11.14 显示了 11 省(市)的城乡权数.

表 11.14

省(市)名	城镇权数	农村权数	省(市)名	城镇权数	农村权数
北京	0.2075094	0.1404956	安徽	0.5561167	3.5888574
江苏	0.4638121	1.5391012	江西	0.5278843	2.184652
河北	0.3130015	2.0316902	青海	0.2276107	0.5781176
广东	0.5439969	2.308378	甘肃	0.1105362	0.5695325
吉林	0.4194826	0.573159	贵州	0.123943	0.8720199
湖北	0.4236392	1.6807483			

对总体域, 我们一方面进行省级、城乡的加权, 同时考虑到男女两性样本回收数, 还分别对性别进行加权. 因每个省内男女两性样本比例近似, 故对性别的加权统一考虑. 11 个省(市)问卷回收总数为 23741 份, 其中男性问卷数 11265 份, 女性问卷为 12476 份. 1990 年第四次人口普查 10% 抽样资料显示, 全国男性人口比例为 51.45%, 女性人口比例为 48.55%, 故男性权数与女性权数分别为:

$$W_{\text{男}} = \frac{23741 \times 0.5145}{11265} = 1.0843092;$$

$$W_{\text{女}} = \frac{23741 \times 0.4855}{12476} = 0.9238742.$$

因此, 计算总体域时, 对各省数据分别用四个权数值修正. 假设某省的省级权数为 $W_{\text{省}}$, 城乡权数分别为 $W_{\text{城}}$ 、 $W_{\text{乡}}$, 那么:

$$W_{\text{城男}} = W_{\text{省}} \times W_{\text{城}} \times 1.0843092;$$

$$W_{\text{城女}} = W_{\text{省}} \times W_{\text{城}} \times 0.9238742;$$

$$W_{\text{乡男}} = W_{\text{省}} \times W_{\text{乡}} \times 1.0843092;$$

$$W_{\text{乡女}} = W_{\text{省}} \times W_{\text{乡}} \times 0.9238742.$$

表 11.15 列出了每个省这四个权数的值

表 11.15

省(市)名	城男权数	城女权数	乡男权数	乡女权数
北 京	0.2255139	0.1921468	0.1523406	0.1298002
江 苏	0.502959	0.4285409	1.7339201	1.4773683
河 北	0.3408409	0.2897283	2.1534628	1.8603964
广 东	0.5868543	0.5021971	2.506255	2.1323365
吉 林	0.4548488	0.3875491	0.6214515	0.5295268
湖 北	0.4593558	0.3913893	1.8224524	1.5527999
安 徽	0.6090024	0.5137818	3.8372156	3.269459
江 西	0.5725898	0.4876986	2.3688332	2.0183436
宁 海	0.2468003	0.2102836	0.6268592	0.53401079
甘 肃	0.1198554	0.1021215	0.6175384	0.5261671
贵 州	0.1343925	0.1145077	0.9455516	0.8056302

评 注

1) 中国妇女地位调查是一项全国性的大规模的社会调查。它采用三种调查方法,以个人问卷调查为主,辅以社区及企事业单位调查与统计文献调查。其中第二种调查是为第一种调查提供进一步的背景、补充与校验资料。在§ 11.6 北京地区专业技术人员现状抽样调查中,也曾对每个被抽中的基层单位进行过类似的调查。至于本文所说的统计文献调查,也即通过查阅现成各种统计资料,获得与调查目标相关的种种信息更是一项完整调查所必不可少的,特别是在撰写调查报告,对所感兴趣的问题进行进一步分析研究以及作预测或其他决策时,尤其如此。

2) 本项调查的组织严密,2000 余名调查员经过系统培训,事先组织了两次试点调查,在调查实施以及数据录入整理各个环节都有严格的质量检验措施,这些都是一项调查取得成功的保证。

3) 本项调查的抽样设计也是十分严密的,首先考虑到今后分析的需要,将总体按省分成研究域(domains of study),省内再按城市与乡村分域。这里的域实际上就是需加以研究的子总体,在抽样时,与层的处理完全相同。设计的基本方法也是采用分层多阶不等概率抽样,在每个域内再按某些特征分层(这仅是为了提高精度),层内采用抽县(市),县(市)内抽

街道(乡), 街道(乡)内抽居(村)民委员会, 居(村)民委员会内抽户的四阶抽样, 以家庭户作为基本抽样单元。在前三阶抽样中都用 PPS 抽样, 最后一阶采用固定样本量(城市中每个居委会抽 20 户, 乡村中每个村民委员会抽 10 户)的简单随机抽样或等距抽样。这样设计的样本是严格自加权的。这样可以大大简化数据处理的工作量。文中在计算抽样比时所用的 Mos (measure of size) 是各阶抽样中衡量抽样单元大小的指标。在本例中取的是所包含的户数。有时也常用人口数代替。

4) 按抽样方案, 每个样本户只调查一人(除非该样本户包含的调查对象, 即 18~64 岁的成年人超过 6)。为保证最终样本中的性别、年龄结构与总体大体一致, 理论上讲应是在所有符合条件的调查对象中随机抽取一人进行调查。但这时的随机如何进行控制而不会流于形式变为任意呢? 本例采用了 L. Kish (1965) 设计的几套选择表。将全部调查表标成 A 式、B₁ 式、B₂ 式、C 式、D 式、E₁ 式、E₂ 式与 F 式 8 种选择表, 比例分别为 1/6, 1/12, 1/12, 1/6, 1/6, 1/12, 1/12 与 1/6, 并按此比例分发至(例如说)每个样本街道(乡)。将样本户中所有符合条件的调查对象按先男后女, 先长后幼的顺序(实际上任何一种确定的顺序都一样)编号, 则实际被调查人即是按随机抽取的选择表上所表明的编号的家庭成员。这样设计即可保证样本户中每个符合条件的成员被抽中的概率都相等。例如若某样本户有 4 名符合条件的成员, 那么排在第一位的成员将在抽到 A, B₁ 两种选择表时抽中, 概率为 $(1/6) + (1/12) = 1/4$; 第二位成员则在抽到 B₂ 及 C 两种选择表时接受调查, 概率也为 1/4; 第三位成员则在抽到 D, E₁ 选择表时接受调查, 第四位成员是在抽到 E₂, F 选择表时接受调查, 概率还是 1/4。这种操作方法虽然比较麻烦, 又要印制 8 种不同的标志, 但因规定具体, 且容易事后检查, 从而能更好地保证抽样的随机性。本例中样本的性别比例、不同年龄段的比例以及不同文化程度的比例均符合 1990 年人口普查全国总体的相应比例即是佐证。

5) 前面已提到由于层内样本是自加权的, 因此对平均数与比例的估计直接可用样本数值, 且方差估计也很简单, 但本例未涉及对比值型目标量的估计, 其实这在多指标调查中是不能回避的。另外层以上各个层次的研究域的估计则可简单的采用分层抽样公式, 从而主要是确定层权。只要权数一定, 其他问题就迎刃而解。

§ 11.9 国家卫生服务总调查^{*)}

为了适应社会主义市场经济体制的形成和发展、政府职能转变和科学决策的进程,促进卫生事业宏观管理水平和决策能力的提高,加强卫生事业发展战略目标及其实施过程的监督、监测和评价,国家卫生部《卫生事业第八个五年计划及2000年规划设想》明确提出:“建立卫生发展、管理目标及其监督评价的指标体系和定期的卫生服务总调查制度以及灵活、及时、准确的综合卫生管理信息系统”,部领导已多次强调要加快综合卫生管理信息系统的建设,“为部领导制定方针政策、健全法制服务,为制定卫生事业发展规划服务,为宏观管理实行监督、监测服务。”为此,部长办公会决定1993年在全国范围内开展国家卫生服务总调查,作为完善国家综合卫生管理信息系统的重要环节,为制定社会卫生计划与政策、卫生管理与评价服务。

卫生服务是一个国家或地区卫生部门为一定的目的合理使用卫生资源向人民群众提供卫生服务的过程。卫生服务的调查研究旨在为卫生事业的宏观管理和科学决策提供客观依据。早在五十年代,美国等西方国家就建立了以连续性的健康询问调查为重点的卫生服务调查研究。七十年代起,英国、加拿大、日本、荷兰等一些发达国家也相继建立了健康询问调查制度。近十多年来,一些发展中国家陆续开展了一次性或重复性的横断面卫生服务抽样调查。我国卫生服务的调查研究起步较晚,但发展速度较快、调查研究的规模较大。自1981年4月中美双方合作在上海县开展卫生服务的调查研究以后,相继有长春市等十多个城市和农村地区开展了卫生服务的抽样调查。1985年以来卫生部有关司局相继在全国范围内开展了城乡医疗卫生服务、民族地区医疗服务、卫生防疫、妇幼卫生、乡镇企业职业卫生需求与对策的调查研究。这些调查研究不仅为卫生事业科学管理和制定卫生事业发展规划提供了重要依据,也积累了比较完整的卫生服务抽样调查的经验。在认真吸取国际、国内卫生服务调查的经验,充分考虑调查研究的科学性、可靠性和可行性的基础上,拟定本调查方案。

一、调查目的

^{*)} 本节正文节选自中华人民共和国卫生部《国家卫生服务总调查方案及调查指导手册》,1993。

国家卫生服务总调查的基本目的是提供人群健康状况及卫生服务需要量、有关卫生服务资源的筹集、分配、结构和卫生服务资源利用及其效率的资料,为卫生事业管理决策提供客观依据。具体目的如下:

1. 通过系统地收集我国不同类型地区居民两周病伤的患病率和慢性病患率、伤残率、因病伤丧失劳动能力程度及其影响因素的资料,反映我国和不同类型地区居民的健康状况、卫生服务需要量和存在的主要卫生问题,分析不同类型地区卫生问题的优先级以及主要的影响因素。

2. 从提供卫生服务的种类、数量和居民实际接受各类卫生服务的程度两个方面系统收集我国不同类型地区居民卫生服务利用的资料,分析和评价我国卫生服务利用的效率和效果以及地区间的差异,确定不同类型地区卫生服务资源利用的现状和存在的问题。

3. 系统地收集我国不同类型地区卫生资源的投入量以及筹集、分配、结构、比例,享受各种医疗保健制度的人数、费用以及因病自付医疗保健费用等资料,分析和评价我国不同类型地区卫生服务资源分配和结构的合理性以及影响因素。

4. 分析和研究我国居民健康状况、卫生服务需要、卫生服务利用及卫生服务资源之间的联系,探讨卫生服务供需的平衡关系,为卫生事业的发展 and 改革、宏观管理和科学决策提供依据。

5. 为深入进行某些疾病病因或医疗预防保健措施等方面的专题研究提供线索。

二、调查对象和调查时间

家庭健康询问调查的对象为全国抽中样本住户的实际人口(凡居住并生活在一起的家庭成员和其他人,或单身居住、生活的,均作为一个住户)。卫生机构调查为抽中“样本地区”[包括样本县(市或市区)、样本乡镇(街道)、样本村(居委会)]的卫生机构和基层卫生组织。

国家卫生服务总调查的调查时间从1993年6月1日开始至6月25日结束。

三、抽样设计

国家卫生服务总调查抽样的原则是经济有效的原则。根据调查目的和调查内容采用多阶段分层整群随机抽样方法抽取“样本地区”和“样本个体”。

第一阶段分层采用多变量分析法综合社会经济、文化教育、卫生保健和人口结构等多个指标为分层标识以县(市或市区)为单位进行分层,将

全国 2400 多个县(市或市区)分为五类地区。根据所要求的样本量按各层占总体的比例随机整群抽取各层的“样本县(市或市区)”共 90 个。

第二阶段分层采用人口数或人均收入为标识,以乡镇(街道)为单位。每个“样本县(市或市区)”按 20% 的比例随机整群抽取乡镇(街道),平均每个县(市或市区)抽取五个乡、镇(街道)为“样本乡镇(街道)”,全国共抽取 450 个。

第三阶段采用人口数或人均收入为标识,以村(居委会)为单位,平均每个“样本乡镇(街道)”整群随机抽取两个“样本村(居委会)”,全国共抽取 900 个村(居委会)。

最终的抽样单位是户,在每个“样本村(居委会)”中随机抽取 60 户,全国共抽取 54000 户。全国平均每户被抽取的概率为 1:5000 (见表 11.16)。

表 11.16 国家卫生服务总调查样本量和抽样概率

单位名称	全国总数	抽样样本量	抽样概率
县/市区	2450	90	1:27
乡镇/街道	70000	450	1:154
村/居委会	100000	900	1:1120
户	280000000	54000	1:5000
人	1200000000	216000	1:5000

四、调查内容

国家卫生服务总调查包括基于“人群”的家庭健康询问调查和基于“机构”的卫生服务调查,两种调查内容各有侧重。(具体项目略)

五、调查方法

国家卫生服务总调查采用一次性横断面抽样调查。

1. 资料收集的方法

基于“人群”的家庭健康询问调查采用入户询问、询问与查阅记录相结合的方法。经培训合格的调查员在对调查户进行摸底调查后深入样本户按调查表的项目对该户所有成员逐一进行询问调查;有关调查项目如确定孕产妇系统保健(产前检查、产后访视等)和儿童系统保健等内容应与保健手册的记录核对。

基于“机构”的卫生服务调查采用文件抄录和实地调查相结合的方法。

法, 常规报告、报表和工作记录已有的指标, 可根据调查表具体的要求抄录; 需要调查的指标由样本县(市或市区)卫生局、被调查卫生机构的统计人员与有关人员配合进行实地调查。

2. 收集资料的人员

家庭健康询问调查, 设调查员和调查指导员。调查员负责入户调查。调查员的挑选由当地的医务人员承担为宜, 在农村挑选乡镇卫生院的医生及部分乡村医生, 在城市挑选地段医院医生。非医务人员由于他们在疾病诊断方面存在困难, 一般不予考虑。一般一个样本乡镇(街道)组织两个调查组, 一个调查组应有2名调查员(一名卫生院医生和一名乡村医生, 平均一个调查组调查60户)。

调查指导员负责调查的组织、指导、检查、及验收工作。调查指导员应是乡镇卫生院及以上卫生机构的医生, 由县(市区)卫生局指定, 每个样本乡镇(街道)应配一名。

“机构”卫生服务调查的调查人员应该是该单位的业务领导和统计人员, 调查时需要与有关业务部门的同志配合。

3. 资料收集的工具

家庭健康询问调查采用: ① 家庭健康调查表; ② 0~5岁儿童健康调查表; ③ 15~49岁已婚育龄妇女健康调查表; ④ 60岁及以上老年人健康调查表。⑤ 两周病伤调查表; ⑥ 1992年住院调查表。

“机构”卫生服务调查采用: ① 全县(市或市区)基本情况调查表; ② 乡镇(街道)卫生机构调查表; ③ 村级(居委会)卫生组织情况调查表; ④ 医院(县及县以上医院、中医院、专科医院、疗养院)情况调查表; ⑤ 卫生防疫机构情况调查表; ⑥ 妇幼保健机构情况调查表。

六、调查实施和质量控制

为了保证调查的顺利展开和调查的质量, 必须对调查的每一个环节实行严格的质量控制。质量控制包括设计阶段(含调查表的设计)的质量控制、调查员的质量控制、调查实施阶段的质量控制和资料整理阶段的质量控制。

1. 调查方案设计、论证和试调查

调查方案的设计必须要科学可行, 指标筛选要慎重, 指标解释要清楚, 各项标准要统一, 在正式确定调查方案前必须经过反复的论证和试调查, 其目的是检验调查设计工作的合理性及可行性, 正式调查前通过试调查使调查员熟悉调查内容, 做到准确、完整地填写调查表格。

2. 调查人员的培训

调查人员的严格挑选和培训是取得准确、可靠资料的不可缺少的前提。培训的要求是：明确调查的目的和意义，了解调查设计的原则和方法，统一指标的含义及填写，得以保证调查工作的质量，明确调查工作的进程等。每一个调查员必须按照统一计划和填表说明的要求执行。人员培训按统一的培训计划、统一培训内容和教材分两级培训。卫生部负责培训省级国家卫生服务总调查管理人员和样本县(市或市区)负责人及师资人员，省督促各样本县(市或市区)培训乡镇(街道)调查指导员和调查员。培训结束后，应对培训效果进行考查，考查合格后才能参加正式调查。

3. 明确调查人员工作职责，建立调查质量核查制度

明确调查人员任务与职责分工是保证调查质量重要因素之一，提高调查人员的责任心和积极性，防止由于分工不清和责任不明造成扯皮现象。调查指导员和调查员必须按照《国家卫生服务总调查调查人员职责及现场工作准则》的要求进行工作。

调查质量的核查制度包括：

① 现场调查中，在每户询问并记录完毕后，调查员都要对填写的内容进行全面的检查，如有疑问应重新询问核实，如有错误要及时改正，有遗漏项目要及时补填。

② 每个乡镇(街道)的调查指导员要对每户的调查表进行核验收收，从正式调查开始后的当晚检查调查表的准确性和完整性，发现错漏项时，要求调查员应在第二天重新询问予以补充更正，认真核实无误后，方可签字验收。

③ 每个县(市区)设立质量考核小组在调查过程中抽查调查质量，调查完成后进行复查考核，家庭健康询问调查的复查考核应在已完成户数中随机抽取5%，观察复核调查与调查结果的符合率；机构卫生服务调查的复核应与有关报表如人员、财务、工作报表等核对，考查其符合率。

④ 卫生部将组织有关省成立质量检查组，分赴各地进行质量考核。

4. 质量要求

① 一致性百分比：用来衡量调查人员调查技术的一致性，要求经过培训后，调查人员调查技术的一致性达到100%。

② 符合率：复查考核中，同户复查与调查结果的符合率除了两周患病有所差异以外，其他项目符合率要求在97%以上。

③ 调查完成率：在出现了三次上门无法调查而放弃该户时，应从候

补户数中按顺序递补,调查完成率应控制在98%以上。

④ 本人回答率:回答应以本人为主,本人不在场时应由熟悉情况的人代替回答;儿童一般由母亲代替回答,育龄妇女最好由本人回答;要求成年人自己回答率不低于70%。

七、数据处理及上报方式

采取分省录入,集中汇总的方式。各调查县(市区)如期将调查表收齐审核无误后,在规定的时间内(1993年7月10日前)上交给各省卫生厅,各省卫生厅验收合格后按卫生部统一编制的程序组织人员进行录入,经检查数据无错误、无遗漏后,在1993年8月底前将软盘报至卫生部卫生统计信息中心。

八、组织领导

国家卫生服务总调查由卫生部统一组织,国家中医药管理局和国家医药管理局参与,组成“国家卫生服务总调查领导小组”,邀请有关专家成立“专家咨询组”和有关人员组成的具体“执行组”,具体负责国家卫生服务总调查的方案设计和论证、组织全国省和县级师资培训、组织调查实施、质量控制、技术指导 and 咨询等工作。

各省、自治区、直辖市卫生厅局相应成立领导小组,负责本省抽样地区的卫生服务调查的领导、组织调查实施、质量控制和资料验收、技术指导和咨询等工作。

样本地区的卫生局应成立相应领导小组,负责领导、组织调查指导员和调查员的培训、组织实施本地区卫生服务的调查和调查表的质量控制工作。

各省、自治区、直辖市和各样本县(市或市区)领导小组人员名单以及参加调查工作的调查人员名单在调查完毕后由各省、自治区、直辖市统一报送卫生部。

全国范围内开展综合性的卫生服务抽样调查在我国尚属首次,是一次新的、艰巨的,但也是一次意义重大的工作。要求各地卫生行政部门要给予高度重视,作好组织、宣传和实施工作,取得当地政府及各界人士的支持,做好群众的宣传、教育和组织工作,以取得群众的理解和密切的配合。

附件 国家卫生服务总调查样本地区和样本个体的抽取方法

一、概述

1. 国家卫生服务总调查抽查的原则是既要兼顾调查设计的科学性即样本地区和样本个体对全国和不同类型地区有足够的代表性, 又不致于过多增加样本量而加大调查的工作量, 即经济有效的原则。

2. 抽样的方法是多阶段分层整群随机抽样法。第一阶段分层是以县(市或市区)为样本地区; 第二阶段分层是以乡镇(街道)为样本地区; 第三阶段分层以村为样本地区; 最后是住户为样本个体。

二、第一阶段分层整群抽样

1. 第一阶段抽样着重解决两个基本问题

一是由于全国各县、市差异极大, 如何确定第一阶段分层的基准; 二是抽样比例, 多大的县、市样本量能经济有效地代表全国和不同类型的地区。

2. 第一阶段分层基准的确定

第一阶段分层的指标是通过专家咨询法和逐步回归法筛选的 10 个与卫生有关的社会经济、文化教育、人口结构和健康指标。10 个指标的主成份分析结果如表 11.17 所示。

表 11.17 主要社会经济和人口动力学指标的主成份因子模型

变 量	主成份 1	主成份 2	主成份 3
第一产业就业率 %	0.82*	-0.49	0.17
14 岁以下人口比例 %	0.80*	0.10	-0.49
文盲率 %	0.69*	0.32	0.22
粗出生率 %	0.69*	0.35	-0.10
粗死亡率 %	0.67*	0.51	0.33
婴儿死亡率 %	0.67*	0.60*	-0.03
人均工农业产值	0.65*	0.53*	0.12
第二产业就业率 %	-0.84*	0.45	-0.10
初中人口比例 %	-0.92*	0.02	-0.04
65 岁以上人口比例 %	-0.10	-0.19	0.93*

从主成份分析中可以看出主成份 1 与绝大多数变量有十分显著的关联, 意义十分明确, 而且代表 10 个变量整体信息的 51.22%, 其值的大小可以综合反映一个地区社会经济、文化教育、人口及其健康的发展。因此, 确定主成份 1 为分层的基准称它为分层因子。

3. 第一阶段的聚类分层

在计算各县、市分层因子的得分后,用 K-Means 聚类分析方法将总体分为组间具有异质性和组内具有同质性的五类地区即五层。聚类分层的结果第一层有 201 个县(市或市区),占整个县(市或市区)的 8.2%;第二层有 650 个县(市或市区),占 26.5%;第三层有 698 个县(市或市区),占 28.5%;第四层有 691 个县(市或市区),占 28.2%;第五层有 212,占 8.6%。

表 11.18 显示了各层因子得分和选择的社会经济等变量的均值,可见各层呈明显的梯度,可以认为,第一层所在的市县,是社会经济、文化教育和卫生事业发展以及人群健康状况好的地区,第二层是比较好的地区,第三层是一般性地区,第四层是比较差,第五层是差的地区。

表 11.18 主要社会经济和人口动力学指标的主成份因子模型

层别	市县数	因子得分		社会经济和人口动力学指标				
		均数	距离	GNP	AEP	ILLIT	CDR	IMR
1	201	-2.4354	3210.28	3330	15.7	19.7	5.1	17.5
2	650	-0.6638	2164.66	895	64.6	23.7	5.7	26.2
3	698	0.0692	1655.00	450	83.5	32.4	6.8	31.4
4	691	0.5776	1264.57	341	88.1	43.6	7.4	49.1
5	212	1.7457	539.61	319	90.0	66.8	11.7	121.4

表 11.19 不同大小样本量样本在各层的分配

层别	全国		不同大小样本量样本的分配				
	合计 (%)		120	90	60	45	30
第一层	201 (8.2)		10	8	5	4	2
第二层	650 (26.5)		32	23	16	11	8
第三层	698 (28.5)		34	26	17	13	9
第四层	691 (28.2)		34	25	17	13	8
第五层	212 (8.6)		10	8	5	4	3

4. 第一阶段分层等概率多种样本量的抽样

用经济有效的样本代表总体是抽样调查的精髓。样本量的确定基于以往的经验和其他国家抽样调查样本的设计,首先给定一个样本量大小

的范围,确定抽取样本量为 120, 90, 60, 45, 30 五个大小不等的样本,为了保证各层每一个县(市或市区)都有同等被抽取为样本的概率,必须考虑不同大小样本量的样本在各层的分配,即按比例的分层抽样,见表 11.19.

按系统随机抽样方法,每个不同大小样本量的样本抽取 6 次. 同样样本量的 6 次抽样,通过计算每次抽样样本各变量的统计量,分别与总体各变量参数进行比较,从中筛选出与总体参数最为接近的那个样本,作为该样本量的最佳抽取样本.

考虑到经济有效的原则和对全国、不同类型的地区和上述每个指标的代表性,国家卫生服务总调查的县(市或市区)样本量取 90.

三、第二阶段整群随机抽样

1. 在上述抽取的 90 个“样本县(市或市区)”中,以乡镇(街道)为第二阶段整群系统随机抽样单位.全国每个乡镇(街道)被抽取为“样本乡镇(街道)”的概率是 1:160. 第二阶段整群系统随机抽样全国共抽取 450 个乡镇(街道). 平均每个“样本县(市或市区)”抽 5 个乡镇(街道). 第二阶段分层整群抽样具体由各样本县(市或市区)按下述方法抽取.

2. 第二阶段整群随机抽样的基准

由于一个县(市或市区)内社会经济、文化教育和卫生状况的差异远小于全国各县、市之间的差异,因而确定县(市或市区)的抽样基准相对容易. 根据我国各县(市或市区)的基本特征、实际的可操作性和以往抽样调查常用的指标,确定采用人口数(或人均收入)作为分层基准.

3. 第二阶段整群随机抽样的方法

① 将样本县(市或市区)所有的乡镇(街道)按人口数的多少(或人均收入的大小)由多到少依次排序.

② 由多到少依次计算人口数(或人均收入)的累计数.

③ 计算抽样间隔,用累计的人口总数(或人均收入累计总数)除以抽取的样本数(累计总数/5).

④ 用纸币法(随便拿出一张人民币,看人民币的号码与最初累计数哪一个数接近,取这个数为开始数)随机确定第一个样本乡镇(街道),然后加上抽样距离确定第二个样本乡镇(街道),依次类推确定第三至五个样本乡镇(街道).

四、第三阶段随机抽样

1. 第三阶段随机抽样的基准和样本量

(1) 在同一个乡镇(街道)内, 各村(居委会)的经济发展和卫生状况基本上变异不大, 因此, 第三阶段不用分层, 直接采用随机整群抽样的方法从“样本乡镇(街道)”中抽取样本村(居委会), 但是, 抽样时应按各村人均收入或人口数作为标识进行排序, 第二阶段随机抽样由调查指导员负责。

(2) 每个“样本乡镇(街道)”整群随机抽取 2 个村(居委会), 全国共抽取 900 个村(居委会), 全国每村(居委会)被抽为样本的概率为 $1:1120$ 。

2. 第三阶段整群随机抽样的方法

① 将样本乡镇(街道)所有的村(居委会)按人均收入的多少(或人口数的大小)由多到少依次排序。

② 由多到少依次计算人均收入(或人口数)的累计数。

③ 计算抽样间隔, 用累计总数除以抽取的样本数(累计总数/2)。

④ 用纸币法(随便拿出一张人民币, 看人民币的号码与最初累计数哪一个数接近, 取这个数为开始数)随机确定第一个样本村(居委会), 然后加上抽样距离确定第二个样本村。

五、样本户的抽样

1. 最终的抽样单位是住户, 在每个“样本村(居委会)”中按 20% 的比例随机抽取住户, 平均每个村抽 60 户, 全国共抽取 54000 户。全国平均每户被抽取为样本的概率为 $54000/28000$ 万, 约五千户中抽一户。如果按每户四个人计算, 人口抽样比也为 $1:5000$ 左右。

2. 抽户方法是各样本乡镇(街道)的调查指导员上述抽样比例在样本村(居委会)随机抽取, 具体方法:

① 按人口普查的编码顺序, 按门牌号、楼号、单元号、门号从小到大排列。

② 对同一门牌号, 同一个大院和楼号的, 按门号从小到大排列, 对同一门牌号内没有门号的按从左到右、从外到里、从下到上的原则编码, 一经编码不许变动。

③ 编好住户码列入住户清单中。

④ 根据抽样比例计算应抽的户数(一般平均每个样本村 60 户), 然后系统随机抽取, 方法同上: 第一步将所有住户的人口累计数、本村的平均人口数($1200/300=4$)和本村应抽取的住户数($300*20\%=60$); 第二步计算抽样距离($1200/60=20$); 第三步确定第一个随机数(如取一张人民币, 其编号的后两位数是 12, 这个随机数接近第 3 编号的累计数, 因此确

定第3号住户为第一个样本;第五步用第3号的累计数加抽样距离($13+20=33$),看33最接近第几编号住户,并确定这家住户为第二个样本,同理用第二个样本住户对应的累计数加抽样距离确定第三个样本,同样确定以后各样本住户。

⑤ 抽样时可多抽取六户,作为备用。抽取方法是在上述抽取完毕以后,按上述步骤再从未抽取的住户中抽取6户。

评 注

1) 本项调查旨在全面了解和掌握我国城乡居民健康状况、卫生服务需求量及卫生资源筹集利用情况,为制定我国卫生事业发展规划、方针和政策提供客观依据。调查涉及面广,实际调查了90个县(市或市区)的卫生机构以及5万户家庭,20多万人。本次调查是多目标的综合调查。单就调查表来说,就有《家庭健康询问调查表》、《(县、市)卫生基本情况调查表》、《乡级卫生机构情况调查表》、《村级基层卫生机构情况调查表》、《医院基本情况调查表》、《卫生防疫机构调查表》及《妇幼保健机构调查表》等7种。而每种少则包含数十个问题,多则有上百个问题。其中家庭健康询问调查表又分:住户健康询问表、0~5岁儿童健康调查表、15~49岁已婚育龄妇女健康表、60岁及以上老年人健康调查表、两周病伤调查表以及1992年(调查前一年)住院调查表等6种。其中住户健康询问表既有针对户主的,又要求每个家庭成员回答的,后两种表也是需要每个成员回答的。这确是名符其实的国家卫生服务总调查。像这类调查,不仅要求调查目的与对象明确,调查方法科学且可操作,更需要强有力的组织领导和严格的质量保证措施。本案例中的前一部分对这些方面都作了详尽的介绍,值得借鉴。

2) 本项目中的主调查——即家庭健康询问调查实际采用的是分层四阶整群抽样。即在对全国所有县、市分层的基础上,在层内抽县、市;在县(市)内抽乡(镇、街道);在乡(镇、街道)内抽村(居)民委员会,最后在抽中的村(居)民委员会中抽家庭户,对所有被抽中的户进行全户及每个成员的调查。文中的“样本地区”和“样本个体”即是每阶抽样中的抽样单元。而前两阶抽样中所谓的“整群”抽样的提法不是标准用法,实际上本项目的整群抽样应是指最后一阶即第四阶抽样是以住户为群的抽样。

3) 第一阶抽样即全国对县市的抽样是按比例分配的分层抽样。我国幅员广大,各地经济文化卫生水平差异极大,因此对县、市的分层是十

分必要的。本案例用多元分析方法先根据筛选出来的10个反映社会经济和人口动力学指标进行主成分分析,以第一主成分作为分层指标,再应用聚类分析法将全国2452个县(市、市区)分成5层,各层之间的社会经济和人口动力学指标都有显著差异,这样做将充分利用分层抽样的优点,大大提高调查精度,在条件允许时是十分可取的方法,而且分类(层)的结果还可为以后其他全国性调查作为参考。

层确定后,本例还对第一阶抽样的不同样本量的选择进行了研究,对每种样本量都模拟抽了6个不同的样本,比较样本(县、市)中各变量参数与全国相应参数,最后考虑到经济而有效的原则以及样本对全国的代表性确定采用抽取90个县、市及最终样本。模拟抽样若干组不同的样本,再进行人为的取舍,削弱了样本的随机性,有点“代表性抽样”或“目的抽样”的意思,这是不是一种可取的方法,不能一概而论。但笔者认为除非不得已(例如§11.2中国5岁以下儿童死亡抽样调查中的情况),还是应首选严格的随机抽样。一般情况下,只要样本量不是太小,所获得的样本的代表性是不成问题的。由于多阶抽样中第一阶抽样的抽样误差在整个抽样误差中占主导地位,因此只要组织及费用有保证,第一阶抽样的样本量还是以适当大一些为宜。

4) 第二、三阶抽样都采用按人口数或人均收入进行排序,并以这两个标识之一为辅助变量进行不等概率系统抽样,样本量是固定的。根据方案,这两个标识是可供选择的(其中第二阶抽样首选的是人口数,第三阶抽样首选的是人均收入)。另外第四阶抽样是按20%的比例用无关标识排队的系统抽样抽家庭户,平均样本量为60户。这里有几个问题值得商榷:按某种标识进行排队再作系统抽样是为了增大样本内方差从而提高估计量的精度,这样做是可以理解的。但作为辅助变量进行不等概率系统抽样(在实施时按这个辅助变量累计作为代码)则用人均收入并无意义。若这两阶抽样中皆用人口数作不等概率系统抽样,而最后一阶抽样是固定样本量(不是固定抽样比!这两者不可兼得)的等概率系统抽样,那么所得的样本在县、市内是自加权的。如果进一步第一阶抽样也采用与人口数成比例的PPS抽样(或不等概率系统抽样),则整个样本是自加权的,这样将大大简化其后的数据处理。按本方案实际抽样所得的样本不是自加权的,其数据处理将十分复杂。笔者未见到本案例完整的数据处理公式。但从方案叙述来看,估计设计者将按各阶(平均)抽样比逐级加权或干脆作为自加权(将按人的总抽样比取为1:5000)。这样做误差太

大,这是本案例抽样方案的不足之处。另外最后一阶抽样的样本量(平均为60户)按笔者的经验也稍为多了一些,如果将第二阶抽样中的街道(乡、镇)或第三阶抽样中的村(居)民委员会的样本量增加一倍,将每个村(居)民委员会的样本量减少一半将可进一步提高精度。

5) 本例各阶抽样实施时均采用随机起点的系统抽样,起点单元(或代码)的抽取用的是纸币的末几位号码。这个方法固然可行,看起来似乎也简单,但实际上存在的问题并不少。首先是“看人民币的号码与最初累计数哪一个数接近,即这个数为开始数”,这与不等概率抽样的累计代码法确定样本单元并不完全相符,因而是严格的。用纸币号码代替随机数的产生看起来是为了避免使用随机数表或随机数骰子,其实实现随机化机制有多种,对于系统抽样中只需产生少数几个随机数的场合,笔者建议利用目前已相当普及的袖珍计算器。对于一般科学计算器,按一下标有(SHIFT) RAN 的那个键(一般是小数点“.”键或数字“0”键的第二功能键)即可产生一个 $[0, 1]$ 范围内的均匀随机数(实际产生的一般是 $0.000 \sim 0.999$)。如果要求产生一个 $1 \sim k$ 范围内的随机整数,则可用 k 乘以计算器产生的随机数,再取整就得到所需要的随机数。读者不妨一试。

§ 11.10 人口变动情况抽样调查^{*}

为保证人口变动情况抽样调查对全国和各省、自治区、直辖市有较好的代表性,本调查以全国为总体,以省级单位为子总体。

一、各省、自治区、直辖市样本量

设计样本规模的主要参数是人口出生率 OBR, (最大)允许绝对误差 Δ , 置信度 $1-\alpha$, 抽样比 f 和设计效应 deff。各省、自治区、直辖市应根据 1992 年人口出生率,采用不同的抽样精度确定各自的样本量,其中 Δ 控制在 $1\% \sim 1.8\%$ 范围内,相对误差控制在 10% 左右,置信度取为 95% ($t=2$), deff 估计为 1.4。样本量的计算公式为:

$$n = \frac{t^2(OBR)(1-OBR)}{\Delta^2} \times \text{deff}.$$

各省级单位拟抽取的样本量和允许抽样误差见表 11.20。全国样本总量为 116.2 万人,人口出生率允许(绝对)误差约为 0.3% 。

^{*} 本节正文根据国家统计局《1993 年人口变动情况抽样调查方案》(1993 年 9 月)及有关附件改写。改写目的是为使原文的表达更清楚些。

表 11-20 1993年人口变动情况抽样调查拟抽取样本量和第一级抽样比

地 区	1992年 出生率 (%)	1992年 总人口 (万人)	样本量 (万人)	允许误差	相对误差	抽样比	1992年 县级单位 个 数	抽取县级 单位比例	抽取县级 单位个数	每县级单 位抽查 小区个数
国 全	18.24	117171	116.2	0.0003	0.02	0.0010	2323	0.25	719	6
京 北	9.22	1102	3	0.0013	0.14	0.0027	18	1.00	18	7
津 北	12.50	920	3	0.0015	0.12	0.0033	18	1.00	18	7
蒙 西	15.33	6275	4	0.0015	0.09	0.0006	173	0.17	80	5
宁 西	19.59	3979	4	0.0016	0.08	0.0013	118	0.25	30	5
林 东	17.07	2207	4	0.0015	0.09	0.0018	100	0.25	25	6
江 南	12.57	4016	4	0.0013	0.10	0.0010	100	0.25	25	6
苏 东	15.74	2532	4	0.0015	0.09	0.0016	59	0.34	20	8
江 南	16.25	3603	4	0.0015	0.09	0.0011	132	0.20	26	6
浙 东	7.28	1345	3	0.0012	0.16	0.0022	20	1.00	20	6
安 南	15.71	6911	4	0.0015	0.09	0.0006	106	0.25	27	6
福 南	14.72	4236	4	0.0014	0.10	0.0009	86	0.25	22	7
江 南	18.76	5334	4	0.0016	0.09	0.0007	103	0.25	26	6
建 南	13.18	3116	4	0.0016	0.09	0.0013	81	0.25	20	8
西 南	19.53	3913	4	0.0016	0.08	0.0010	99	0.25	25	6
东 南	11.43	8610	5	0.0011	0.10	0.0006	135	0.22	30	7
北 南	18.13	8331	5	0.0014	0.08	0.0006	157	0.20	31	6
南 南	19.05	5330	4	0.0016	0.08	0.0007	98	0.25	25	7
湖 南	16.70	6267	4	0.0015	0.09	0.0006	122	0.25	30	5
湖 东	19.31	6525	4	0.0016	0.08	0.0006	119	0.25	30	5
湖 西	20.19	4380	4	0.0017	0.08	0.0009	104	0.25	26	6
海 南	21.31	636	4	0.0017	0.08	0.0009	20	1.00	20	8
四 南	16.27	10998	5	0.0013	0.08	0.0005	219	0.14	30	7
贵 南	22.40	3361	4	0.0018	0.08	0.0012	86	0.25	22	7
云 南	21.00	3322	4	0.0017	0.08	0.0010	127	0.20	25	6
西 南	23.63	223	0.2	0.0080	0.34	0.0009	78	0.05	4	2
陕 西	18.85	3405	4	0.0016	0.09	0.0012	107	0.25	27	6
海 南	19.37	2314	4	0.0016	0.08	0.0017	85	0.25	21	8
夏 南	22.54	461	4	0.0018	0.08	0.0087	43	0.47	20	8
宁 南	20.11	487	4	0.0017	0.08	0.0082	24	1.00	24	7
新 疆	22.80	1531	4	0.0018	0.08	0.0025	96	0.25	24	7

二、抽样方法

全国多数省级单位采用分层三级整群,与人口数成比例的概率抽样,直辖市和个别省则采用分层二级整群,与人口数成比例的概率抽样。

1. 抽样框

三级抽样框中的抽样单元分别为:县(市、区),乡(镇、街道)与调查小区[村(居)民小组或自然村]。各省、自治区、直辖市可根据 1990 年人口普查行政区划资料和人口数建立各级抽样框,可按实际情况进行调整,但应保证不重不漏。第三级抽样单元的调查小区可以是村(居)民小组或自然村,人数控制在 250 人左右。同时调查小区必须是一个完整的地域。

2. 分层原则和方法

各省、自治区、直辖市要对一级抽样单元县(市、区)级单位进行分层。分层原则应尽可能使层内各单位之间人口变动指标差异减少,各层之间差异增大。各地区应根据 1990 年人口普查和 1992 年人口变动情况确定分层标志。由于各地区人口变动情况同本地区农业、非农业人口比例,经济文化发展以及地理条件相关,故县级单位可按人口变动情况、经济标志(经济发达县、经济不发达县)或地形标志(山区、平原、丘陵)等分层。

第一级抽样单元分层后,要按全省的抽样比和各层总人数分配每个层调查的样本量。

3. 各级抽样抽取样本单元数的确定

省级单位中第一级抽样比 f_1 拟定为 25% 左右(详见表 11-20)。第二级抽样抽取的乡级单位个数根据层内调查小区的平均规模确定。原则上在每个层抽中的县级单位内应抽 3~4 个乡(镇、街道);每个乡级单位抽 2 个调查小区。

4. 各级抽样方法

第一级抽样:层内县级单位按 1992 年人口出生率高低或其他有关标识排队,并按排列的序号将各单位人口累计,在人口累计栏中,随机等距抽取县级单位。

第二级抽样:在被抽中的县级单位内,将各乡级单位也按与人口出生率高低有关的标识排队。在排列乡级单位时,应将乡、镇、街道分类排列,进行隐含分层。并按排列序号将各单位人口累计。在人口累计栏中,随机等距地确定所抽取的乡级单位。

第三级抽样:在被抽中的乡级单位内,各调查小区按地址码排队,用等距抽样抽取所需要的调查小区,然后调查整个小区的人口。

三、人口出生率、死亡率、自然增长率和总人口数的估计

各省、自治区、直辖市由调查样本估计本地区人口变动主要指标时, 人口出生率、死亡率、自然增长率均采用样本平均值. 具体计算公式如下:

$$\text{人口出生率: } \hat{C}\hat{B}R = \frac{\text{调查年出生人数 } y_B}{\text{年平均调查人数 } x}, \quad (11.87)$$

$$\text{人口死亡率: } \hat{C}\hat{D}R = \frac{\text{调查年死亡人数 } y_D}{\text{年平均调查人数 } x}, \quad (11.88)$$

$$\text{人口自然增长率 } \hat{R} = \hat{C}\hat{B}R - \hat{C}\hat{D}R, \quad (11.89)$$

$$1993 \text{ 年底人口总数 } \hat{N} = \frac{2 + \hat{R}}{2 - \hat{R}} N_0 \quad (N_0 \text{ 为 } 1992 \text{ 年底人口数}). \quad (11.90)$$

四、省级人口变动情况抽样误差计算公式

以省级人口出生率 CBR 的估计 \hat{R} 为例, 其方差估计 $v(\hat{R})$ 按以下公式计算:

若 y_{hi} 与 x_{hi} 分别是 h 层 i 县(市)样本出生人数及年平均人数(调查点为年初人口与年末人口的平均数), L 是全省层数, n_h 是 h 层调查的县(市、市区)数, 则

$$\hat{R} = \frac{y}{x} = \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}}, \quad (11.91)$$

$$v(\hat{R}) = \frac{1}{x^2} [v(y) + \hat{R}^2 v(x) - 2\hat{R} \text{Cov}(x, y)], \quad (11.92)$$

其中

$$v(y) = \sum_{h=1}^L n_h S_{y_h}^2 = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2, \quad (11.93)$$

$$v(x) = \sum_{h=1}^L n_h S_{x_h}^2 = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2, \quad (11.94)$$

$$\text{Cov}(x, y) = \sum_{h=1}^L n_h S_{xy_h}^2 = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h). \quad (11.95)$$

而 \bar{y}_h 与 \bar{x}_h 分别是 h 层中样本县(市)中的平均出生人数与调查人数.

在 95% 的置信度下, \hat{R} 的允许绝对误差 d 与相对误差 r 分别为

$$d = 1.96 \sqrt{v(\hat{R})},$$

$$r = d/\hat{R}.$$

评 注

1) 人口变动情况抽样调查是人口统计的一个重要组成部分, 全国人口普查只能间隔许多年(我国规定今后每隔 10 年)进行一次。在其他年份为及时摸清人口变动情况只有搞抽样调查。我国自 1982 年第三次人口普查后, 每年 1 月 1 日进行全国人口变动情况抽样调查, 调查前一年 1 月 1 日零时至 12 月 31 日 24 时人口出生、死亡、迁入与迁出情况, 进而推算年底时的人口总数。此项调查每年为国家提供可靠的人口信息资料, 供制定国民经济和社会发展计划, 制定人口政策服务。十多年来, 此项调查的抽样方案几经演变, 逐渐完善。上面综述的是 1994 年 1 月 1 日执行的为调查 1993 年间人口变动情况的最新方案。

根据本方案, 全国每个省级单位(包括自治区及直辖市)都需进行调查, 而且分别进行数据处理。整个设计是基于调查不仅对全国而且对每个省级单位都有代表性意义的基础上的。全国所有调查点(调查小区)的人口数约为 116 万人, 总的抽样比约为 1%。对全国的设计精度, 按人口出生率的绝对误差限为 0.3%(置信度 95%), 规模是比较大的, 精度也比较高。调查所得的数据应是合理可靠的。

2) 各省级单位的抽样多数采用了分层三阶整群抽样, 先将省内各县(市、市区)分层, 再按县(市), 乡(镇、街道)与调查小区(村或居民小组)三阶抽样, 对抽中的调查小区进行全面调查。对直辖市及海南、宁夏则省掉对县(市、市区)这一阶的抽样。这种尽可能降低抽样的阶数以及缩小群的规模对减小抽样误差、提高效率都是有好处的。另外, 本例中前两阶抽样都采用按人口数成比例的不等概率系统抽样, 而单元排列按人口出生率或其他有关标识排队, 这样做都是为了提高估计精度。事实上, 本方案中取的设计效应 $deff$ 仅为 1.4, 是比较低的。顺便提一句, 由于此项调查是定期进行的, 因此 $deff$ 可从以前的调查中获得较精确的估计。

3) 样本量的分配, 除西藏外, 每个省级单位抽取 3~5 万人, 其中三个人口最多的省——四川、河南、山东抽取略多, 为 5 万人; 三个直辖市由于总人口数不是太多, 又只是二阶抽样, 故抽 3 万人, 其他省(自治区)都是 4 万人。这与要求每个省级单位最后结果都有意义、都有精度要求这一点是吻合的。西藏自治区由于调查过于困难, 而且人口数又最少, 故对它放宽要求, 仅抽 5% 的县级单位, 共 2000 人, 这些考虑都是从实际出发的。

省级单位内第一阶抽县(市、市区)在各层中系按比例分配。第一、二阶抽样都是按人口比例的不等概率抽样,其中第二阶抽样的样本量即县(市、市区)内抽乡(镇、街道)数根据层内调查小区的平均规模确定,原则上在每层抽中的县级单位内应抽3~4个乡级单位。第三阶样本量是固定的2个调查小区,小区的平均规模是250人左右。这里第二阶抽样的样本量未确定。确切地说:县级单位中所抽的乡级单位数应与当地的调查小区的平均规模成反比,即若调查小区规模小,就应多抽乡级单位。在此条件下,所得的样本是自加权的。由于这条件在实际中不易做到,调查小区即使在一个县级单位内规模也不可能做到完全相等,因此,根据本方案所得的样本在省内只可能是近似自加权的。

4) 按自加权样本处理数据,各省出生率、死亡率与人口增长率及人口总数的估计公式(11.87)~(11.91)都是正确的。方差估计 $v(\hat{R})$ 的公式(11.92)也是对的。公式(11.93)~(11.95)是考虑到层内各县市的抽样是相互独立的,因此(例如对公式(11.93)),

$$\begin{aligned} v(y) &= v\left(\sum_{h=1}^L \sum_{i=1}^{n_h} y_{hi}\right) = \sum_{h=1}^L v\left(\sum_{i=1}^{n_h} y_{hi}\right) \\ &\approx \sum_{h=1}^L n_h v(y_{hi}) = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2. \end{aligned}$$

5) 本案例中的前两阶抽样都是系统抽样,而实际上系统抽样的方差估计是相当困难的。本例是将样本简化成自加权样本来处理的。如果要更精确地进行处理,则应改变前两阶抽样方法,或用放回PPS抽样,或将层再缩小,在每层中按Brewer、Durbin等方法抽取2个单元,在这种情况下,就可以精确地估计目标量及其方差。

§ 11.11 农村抽样调查网点抽选方案^{*)}

一、制订目的

国务院1983年8月批转的国家统计局《关于加强农村统计工作等问题的报告》,要求加强农村抽样调查工作,并批准增加全国农村抽样调查队的编制人数。为了更好地发挥调查队的作用,提高农村抽样调查资料的质量,决定将原有农产量抽样调查和农经抽样调查两套调查县合并,总

^{*)} 本书正文节选自国家统计局:《农村抽样调查网点抽选方案(试行)》,1984,见国家统计局农村社会经济调查总队编:《农村抽样调查基础工作规程》,北京科学技术出版社,86~44。

规模适当扩大,县以下两套网点分别抽选。为此,制订本方案,以便按照方案要求和规定方法,抽选确定农村抽样调查网点,开展调查工作。

二、抽样范围

凡是国家统计局统一要求组织的农产量、农村住户和农村经济抽样调查,应以省、自治区、直辖市为范围严格按照本方案提出的抽样原则、抽样方法、抽样数目,抽选确定调查网点,以保证调查工作的科学性和可靠性。

三、调查内容

在本方案调查网点进行调查的内容,主要有以下各项:

(1) 农产量抽样调查: 各个主要农事季节农作物(当前主要调查粮食作物)的面积和预计、实测产量。

(2) 农村住户调查: 农民家庭的生产、收入、分配、积累、消费、出售和购入商品以及有关社会情况等调查。

(3) 农村经济基本情况调查: 根据国家研究制定政策、编制计划的需要,组织的一次性专题调查。

四、抽样原则

农村抽样调查网点的抽选,必须坚持随机原则,按照科学的抽样方法抽选出的调查单位,应对农产量调查、农村住户调查和农村经济基本情况调查,特别是粮食产量、农民收入都具有充分的代表性。抽选出来的农产量调查的村以上调查单位,农村住户调查全部网点(包括调查户),基本固定,连续观察。

五、抽样方法和抽样数目

根据我国农村和农业经济的具体情况,以及各级党政领导的需要,农村抽样调查网点的抽选,采用多阶段、随机起点、对称等距抽样方法,一般分为省抽县、县抽乡、乡抽村(即自然村、村民小组,下同)、村抽地块或农户等几个阶段进行。各阶段抽样方法如下:

1. 省抽县

(1) 抽样数目: 各省、自治区、直辖市应抽调查县数合计,应占全国总县数的35%左右。县数较少的省、自治区、直辖市可大于这一比例;县数较多的省、自治区可小于这一比例。

(2) 抽样方法: 将经过加工整理的全省各县(即总体各单位)的有关标识和辅助资料,按高低顺序排队,编制排队表(即抽样框)。排队标识和辅助资料有下列两种:

① 近三年平均每公顷粮食产量由低到高顺序排队,以粮食作物播种面积为辅助资料,逐单位累计,按规定县数计算抽样距离。

② 近三年平均每人从集体分配收入由低到高顺序排队,以参加分配人口为辅助资料,逐单位依次累计,按规定县数计算抽样距离。

这两种标识的选择,要根据差异程度大小而定。哪一种差异程度大,即以哪一种资料作为排队标识。排队表(抽样框)编制完成后,即进行抽样。抽选时,先计算抽样距离,然后按对称等距抽样方法的要求,以随机起点(第一组距内的任何一点)开始,按照计算的距离和样本单位位置,抽选出各调查县。

2. 县以下调查网点的抽选

根据调查内容和要求不同,农产量和农村住户调查分别进行。

(1) 农产量调查

① 抽样阶段:农产量调查要求在抽中调查县抽选村进行调查。具体抽选划分阶段,可以先抽乡,从抽中乡抽选村;有条件的也可以由县直接抽选村。

② 抽样数目:

A. 县抽乡、乡再抽村的县,每个调查县一般应抽6至10个乡。每个调查乡应抽的村数可根据每县抽乡的数量多少而决定,一般每乡应为3至5个村,从而保证每县共抽18至30个村。

B. 县直接抽取村的,一般可抽15至20个村。县直接抽行政村(即村民委员会,下同)的,可抽8至12个行政村;行政村再抽村,一般抽3个。

③ 抽样方法:

A. 县抽乡时,应将全县各乡的近三年每公顷的粮食平均产量作为有关标识,按高低顺序排队,以近三年粮食平均播种面积作为辅助资料,按排队顺序依次累计,制成排队表(抽样框)。然后按规定抽样数目,以对称等距抽样方法抽选确定调查乡。

B. 乡抽村的排队标识和辅助资料以及抽选方法,与县抽乡相同。

C. 有条件的地区,由县直接抽村,或由县抽行政村,再抽村时所用排队资料及抽样方法,也与县抽乡相同。

④ 村内农产量调查内容与抽样方式:

A. 粮食播种面积:在抽中村全村范围内进行调查。

B. 粮食预计产量:每个季节调查时,在抽中村内核实全部粮食播种

面积,查清种植粮食作物的田块,然后逐块进行估产;或从调查地块中随机抽取 10 至 15 个小面积样本用查棵数粒方法估算产量。

C. 粮食实测产量:将抽中村的全部粮食作物地块,按调查前的预计每公顷产量高低排队,等距抽选部分地块:北方地块面积较大的,每个村至少抽 7 个地块,南方地块面积较小的,每个村至少抽 15 块。在每个调查地块内,按简单等距抽样方法抽 5 至 10 个样本进行实割实测。

D. 在村内抽选农户进行粮食产量调查的,必须取得地块的预计、实测产量资料。具体抽户办法是:将每户全部调查作物地块估产,然后分户计算每户综合单产,按单产高低排队,以每户面积合计数累计,采用随机起点对称等距抽样方法抽取 10 户,在调查户的全部地块进行调查。在调查户的地块上进行实割实测调查时,可以整块单收单打,也可以用简单等距抽样方法每户抽 5 个以上样本进行实测。

E. 村抽地块进行农产量调查的,调查地块不固定,于每季节调查时临时抽选。村再抽取农户,在农户的地块中进行农产量调查的,调查户也不要固定。

(2) 农村住户调查

① 抽样阶段:县以下一般分两阶段进行,即调查县抽村,村抽调查户。如县内村数较多、地域分布广,不宜直接抽村的,可以实行三阶段抽样,即县抽乡、乡抽村、村抽调查户。

② 抽样数目:根据人口比例,每个调查县的调查户数为 90 户至 100 户。调查 70 户以下的县,每个村调查 5 户;调查 80 户以上的县,每个村可以调查 10 户。按上述要求,县直接抽村的调查县,每县应抽 6~14 个村;县抽乡,乡抽村的调查县,每县应抽 3~7 个乡,每个乡保证抽 2 个村。

③ 抽样方法:

A. 县抽村或县抽乡,乡再抽村时,用近三年平均每人分配收入作为有关标识,按高低顺序排队,再以近三年平均分配人口作为辅助资料进行累计,计算抽样距离,采用随机起点对称等距抽样方法,抽选确定乡或村。具体抽选方法与省抽县相同。

B. 村抽选调查户,用全村各户的上年人均生产性纯收入排队。全村各户人均生产性纯收入要采用一次普查取得,它包含家庭经营得到的纯收入、集体经营分得收入、联合体经营分得收入等三个部分,用全户人口计算人均生产性纯收入。村抽调查户时,不用辅助资料(人口数)计算组距,而用规定调查户数除全村(组、队)户数计算出组距。采用随机起点

对称等距抽样方法,抽选确定调查户。

(3) 农村经济基本情况调查

农村经济基本情况调查就在农村住户调查所在的各抽中调查乡和村中进行调查。

六、计算抽样误差

(1) 按每公顷粮食产量或按收入水平排队、分层(分层数按样本单位数)计算层方差、层误差,其公式为:

$$\sigma_i^2 = \frac{\sum (x - \bar{x})^2 f}{\sum f},$$

$$\mu_i^2 = \frac{\sigma_i^2}{n_i} \left(1 - \frac{n_i}{N_i}\right).$$

(2) 利用分层误差计算各阶段误差:

一阶段误差公式:

$$\mu_1^2 = \frac{1}{N_1^2} \sum N_{1i}^2 \mu_i^2.$$

二阶段误差 μ_2^2 、三阶段误差 μ_3^2 的公式同 μ_1^2 。

(3) 计算多阶段误差:

$$\mu_0^2 = \mu_1^2 + \frac{\mu_2^2}{n} + \frac{\mu_3^2}{n \cdot m} + \frac{\mu_4^2}{n \cdot m \cdot k} + \dots,$$

其中: n 为第一阶段抽出第一阶段样本单位数;

m 为第二阶段, 平均每个第一阶段样本单位抽出的第二阶段样本单位数;

k 为第三阶段, 平均每个第二阶段样本单位抽出的第三阶段样本单位数。

在实际计算中可以看到, 第三阶段以后, 虽然还有误差, 但由于数量很少, 影响已不大, 所以可以省略不算。因而计算全省抽样误差, 可以只计算省抽县, 县抽乡, 乡抽村三级, 要求误差系数控制为: 产量在 2% 以内, 收入在 3% 以内。各县参照上式可以计算本县的抽样误差, 计算收入调查的全县抽样误差, 按县抽乡, 乡抽村, 村抽户三段计算, 计算产量调查的全县抽样误差, 只计算县抽乡, 乡抽村两阶段即可。

(4) 计算抽样误差系数:

$2 \times \sqrt{\mu_0^2}$ / 全省平均每公顷产量(或全省平均收入水平)。

七、代表性检查

(1) 省抽县: 调查县抽出以后, 必须进行代表性检查, 检查的方法

是:以抽中调查县的平均标识值与总体相应标识值进行比较,单产水平出入不超过2%,收入水平不超过3%的为有代表性。

(2) 县抽乡,乡抽村:检查方法与省抽县相同。

(3) 代表性检查所用的资料,应为同一口径的全面统计资料。

八、资料整理

农产量调查资料的推算方法用简单算术平均数法。以各调查单位的调查结果推算总体调查结果时,按方案规定方法进行。

农村住户调查和农村经济基本情况调查资料的汇总整理,按方案规定进行。

九、各级调查单位的抽选、确定、变动、调整的审批程序(略)

评 注

1) 我国农村抽样调查(包括农产量抽样调查、农村住户调查与农村经济基本情况调查等二项经常性的调查以及其他一次性的专项调查)是由国家统计局农村社会经济调查总队及其各省队负责操作的。对每个省、自治区、直辖市)用的都是一套网点,即按本方案抽取的县、(乡)、村、农户(或地块)几级网点。抽选出来的农产量调查的村以上调查单位,农村住户调查全部网点(包括调查户)基本固定,连续观察。之所以采用固定网点的办法,是因为实施调查的农调队的建制是固定的,这固然带来不少实施上的便利条件,也由此造成样本难以轮换的缺点。

2) 本方案中的各阶抽样均采用按有关标识排队的随机起点对称等距抽样,首先将各阶抽样单元(县、村等)按1981~1983三年间的平均每公顷产量(或每人从集体分配收入)由低到高的顺序排队,以粮食作物播种面积(或参加分配的人口数)为辅助资料进行不等概率对称系统抽样。正如我们在对前几个案例评注中曾经指出的那样,按有关标识排队,然后采用对称系统(等距)抽样将大大提高估计量的精度,减少抽样误差;而按一定的辅助变量进行不等概率抽样不仅可进一步提高精度,也保证了最后获得的样本是自加权的(在一系列条件保证下),从而简化数据处理工作。在这两点上,本方案是相当成功的,也是它的最显著的特点。

3) 但本方案中各阶抽样样本量的确定都有一定程度的弹性(允许在一定范围内选取),因而实际上不能保证最后获得的样本是自加权的。更为严重的是:由于调查样本是一次确定,连续多年观察的,而在这些年中,由于各单元作物播种面积(或参加分配的人口数)都会有相当程度的变

化,而这些变化不可能是一致的(特别对播种面积),因此按本方案规定的数据处理(简单算术平均法)必然会带来偏倚。其偏倚的程度随着使用时间会愈来愈大。另一方面,鉴于本方案所用的抽样是按有关标识排队的不等概率对称系统抽样,抽样误差的估计十分复杂,不能按本方案给出的一般多阶(等概率)抽样的公式。不过由于播种面积的变化带来的估计量偏倚可采用依赖于抽样时(即1981~1983年间)的播种面积的Horvitz-Thompson估计。尽管随着时间的推移,抽样误差会愈来愈大,但这样做可以使估计量基本上保持无偏。至于按本方案获得估计量的实际方差,则可用第8章或第9章中所述的方法计算。不过无论用哪一种方法,都将是十分复杂的。

4) 连续多年使用同一套调查网点,除了有前面指出的缺陷外,还会使样本逐渐“疲劳”,从而严重地影响调查的质量,增加调查误差,为此国家统计局在1989年又提出《农村抽样调查样本轮换方案(试行)》。鉴于农调队的建制的原因,样本轮换不对县进行轮换,只对村、户(或乡、村、户)逐级实行轮换。从1990年起用新抽选的调查户替代原有的全部调查户。然后每年轮换一次,每次轮换调查户的25%左右,4年内轮换一遍。新样本的抽样方法大体上与原来的网点抽选方案相同。不同点是原方案农产量调查与农村住户调查在县之下是两套网点,而新方案则到村为止都采用同一套网点。在县内抽村时采用先抽选一套大样本的调查村,然后在大大样本调查村内,以同样的抽选方法抽选小样本的调查村。目标量估计仍采用样本平均数,抽样误差也仍采用二阶(或三阶)抽样,每阶都是简单随机抽样的公式,因此前面指出的问题依然存在。在这方面我国农村抽样调查还有不少需要改进的地方。

§ 11.12 人体测量抽样方案^{*}

1985~1987年间,我们为国家标准局、中国服装工业总公司与中国人民解放军总后勤部军需装备研究所等单位,设计了几个有关人的体型尺寸测量的抽样方案[1]、[2]、[3]。在这些抽样调查方案中,所需估计的目标量均不是一般抽样调查项目中所遇到的,在许多文献中经过充分研究讨论的那些总体参数,例如总体总和、平均数、比例或两个总数之比值等。例

^{*} 本节正文节选自冯士雍、孙山泽、毕健《人体测量抽样方案目标量的估计及样本量的确定》,原载《应用概率统计》,1989,第5卷第4期,350~357。

如在为制定服装号型系列标准为目的的抽样, 人体某些尺寸的平均数, 如平均身高、平均胸围或腰围就不具有特别重要的意义。即使这些量的估计能精确到 0.001 mm , 与制定服装号型系列标准, 使它能满足多数人的需要并无直接联系。在这些问题中, 我们更感兴趣的是给出人体的各种尺寸的分布情况, 需要估计的目标量主要以这些尺寸的分位数 X_p 的形式出现。

为达到上述目的, 考虑到我国人口分布的现状以及人体测量的特点, 在制定抽样方案时我们对所考虑的总体进行必要的划分。对每个子总体(例如成年男子, 成年女子、少男、少女; 将校级军官、尉级军官与士兵等)都采用分层整群抽样(但针对不同子总体的情况方案都不尽相同, 例如对男士兵也采用了系统抽样)。我们对不同的调查方案, 考虑了实际任务的需要, 给出对分位数估计精度的提法, 研究了精度与样本量之间的关系。现将制定这些方案时, 对上述问题的各种考虑及解决方法综合报道如下:

一、层的划分及群的组成

采用分层整群抽样是由于人体测量工作本身的特点决定的。制定一个效率高的人体测量抽样方案, 必须考虑到影响人的体形尺寸各个方面的特点, 诸如地域、年龄、职业等的影响, 同时考虑到测量工作的方便。对此, 我们作了以下的处理:

1. 按地域分层

中国疆土辽阔, 人口众多, 且传统地居住稳定, 人员流动较少。多种历史资料表明, 中国人人体尺寸与地域的关系极为密切。我们参考了有关资料, 按人类学的观点将除台湾以外的全国各省、市、自治区分成六个自然区域。在同一自然区域中, 有的由于地理、气候、遗传等因素的影响, 差别仍较大。因此在有些方案(例如[1])中, 我们再进一步根据几种历史资料中各省成年人平均身高的资料, 划分为高、中上、中下及矮四档。因此最终全国各省、市、自治区被划分成 12 个层, 如表 11.21 所示。抽样时按工作方便, 在层内选取一个或几个省、市、自治区进行测量。而为了今后数据分析的方便, 例如能采用样本分位数估计总体分位数等, 在各层中采用按人口总数比例分配的方法。

2. 群的组成

由于人体尺寸测量是件技术性较强的工作, 同时测量的项目也较多, 例如项目[1], 多达 74 项, 为使工作方便, 我们一般的在层内采用随机整

表 11.21 中国人体型的地域划分

自然区域 平均身高	I	II	III	IV	V	VI
矮				湖南 江西	广东 广西	四川 贵州
中下		日 本 青 海	浙江 安徽	湖北	福建	云南
中上		陕西, 宁夏 山西, 河南 西藏	江苏 上海			
高	黑龙江, 吉林 辽宁, 内蒙 河北, 北京 天津, 山东	新 疆				

群抽样。在群的抽取过程中, 特别要注意的是群内个体的年龄结构。资料表明, 不同年龄段的人体型尺寸有明显的差异。但考虑到实地测量的方便, 不可能做到分年龄段调查测量, 否则, 很难保证抽样的随机性。因此抽样方案规定: 整群样本应是一个自然的群体单位, 如一个独立的实际单位, 或一个单位中的一个或几个车间或班组。人数恰好达到方案规定的群体大小(允许有几个人误差)。避免在一个较大单位中人为挑选被测人员, 或听任自流愿意测试的人才测, 以凑够规定的群体大小。这样做的目的是尽量使被测样本中各年龄段的结构与总体中相应结构基本一致。必要时可通过适当选择样本群以调整样本中的年龄结构。例如当中、老年人的被测人数不足时, 可有意选择一些历史较长、老同志较多的单位, 例如多抽一些办公室、科研单位等。上面提到人体尺寸与地域的关系极为密切, 这就涉及被测人员的籍贯问题, 由于这个问题本身比较复杂, 且因为我们测量主要不是从人类学或遗传学观点进行研究, 因此我们对于被测人员的籍贯问题不予考虑(仅加以记录)。允许有非本省籍的人员, 但有一种情况必须排除, 即当某单位是从不属于本层的外地迁移来时, 则不能选作为样本群。

至于职业或工种对人体体形尺寸的影响, 试调查时发现, 从事不同职业人员的体形尺寸并无明显的差异。因此我们的抽样方案只规定不抽测对体形尺寸有特殊要求的单位。对不同行业与职业则不作区分。

3. 关于群的大小

众所周知, 整群抽样的设计效应 $\text{deff}(\text{design effect})$ 为

$$\text{deff} \approx 1 + (\bar{M} - 1)\rho, \quad (11.96)$$

其中 \bar{M} 是平均群体大小, ρ 是群内相关系数. 关于后者, 我们根据四川省试测数据计算得到的群间均方 s_b^2 与群内均方 s_w^2 (均按通常的方差分析表计算) 以及当时实测的平均群体大小 \bar{M} , 由下式即可估计 ρ 值:

$$\hat{\rho} = \frac{s_b^2 - s_w^2}{s_b^2 + (\bar{M} - 1)s_w^2}. \quad (11.97)$$

试测时的 $\bar{M} = 124$, 计算结果为 $\hat{\rho} = 0.00775$. 为提高效率, 同时也为测量的方便, 减少因测试人员疲劳引起测量误差的增大, 我们取 $\bar{M} = 80$, 即一个测量组一天的工作量.

二、分位数估计量的相对精度及具体估计方法

1. 总体分位数估计量精度的提法

前已指出, 我们估计的目标量为总体的各种尺寸的分位数. 总体某个尺寸 x 的 p 分位数 x_p ($0 < p < 1$) 即是满足下式的量:

$$P\{x \leq x_p\} = p. \quad (11.98)$$

应用中较重要的分位数有 $x_{.025}$, $x_{.05}$, $x_{.10}$, $x_{.20}$, $x_{.50}$, $x_{.80}$, $x_{.90}$, $x_{.95}$ 及 $x_{.975}$ 等, 其中 $x_{.50}$ 即是中位数.

根据所测样本, 可按一定方法对 x_p 进行估计, 记估计值为 \hat{x}_p . 对一般的估计量 $\hat{\theta}$, 精度的提法有绝对精度与相对精度两种. 这两种精度都是在一定的概率意义以下, 例如对于给定的置信度 95%, 绝对精度 Δ 即是满足下式的量:

$$P\{|\hat{\theta} - \theta| \leq \Delta\} = 0.95. \quad (11.99)$$

而通常的相对精度, 即是指满足下式的 r :

$$P\left\{\frac{|\hat{\theta} - \theta|}{\theta} \leq r\right\} = 0.95. \quad (11.100)$$

以上两式中的 θ 都是被估计参数的真值. 按这两种定义不易确定精度与样本量的关系, 而且也不一定满足我们的实际需要. 我们提出分位数估计量 \hat{x}_p 的精度定义如下: 对一个很小的数 d (例如 1%), 使满足

$$P\{x_{p-d} \leq \hat{x}_p \leq x_{p+d}\} = 0.95, \quad (11.101)$$

以下我们称 d 为 \hat{x}_p 的 (相对) 精度. 请注意与 (11.100) 式中的 r 相区别.

2. x_p 的估计方法

按照 (11.101) 的定义, 对于同一样本量, 对不同的 p 值, 分位数 \hat{x}_p 所

能达到的实际精度不同(详见下段中的讨论). 此外, 精度还因 x_p 的估计方法的不同而有差异. 由于人体体形尺寸近似遵从正态分布, 因此在估计总体分位数时有两种方法可以采用: 一种是用样本分位数 \tilde{x}_p 估计 x_p ; 另一种是先从样本计算平均数 \bar{x} 及样本标准差 s , 以

$$x'_p = \bar{x} + u_p s \quad (11.102)$$

估计 x_p , 其中 u_p 为标准正态分布 p 的分位数.

由于总体指标的实际分布与正态分布一般有一定差异, 特别是在分布的两端. 此外理论计算表明(详见下段), 对于较小的 p 值(例如 $p \leq 0.2$) 或较大的 p 值(例如 $p \geq 0.8$), 以第一种估计方法精度较高, 而对中间的 p 值($0.2 < p < 0.8$), 以第二种估计方法精度较高. 因此我们采用 x_p 的估计量 \hat{x}_p 为:

$$\hat{x}_p = \begin{cases} \bar{x}, & \text{当 } p = 0.5; \\ \tilde{x}_p, & \text{当 } 0 < p \leq 0.2 \text{ 或 } 0.8 \leq p < 1; \\ \omega(p)\tilde{x}_p + [1 - \omega(p)]x'_p, & \text{当 } 0.2 < p < 0.5 \text{ 或 } 0.5 < p < 1. \end{cases} \quad (11.103)$$

其中 $\omega(p)$ 是适当选取的权, 可与 p 有关($0 < \omega(p) < 1$). 当 p 值接近于 0.2 或 0.8 时, 取 $\omega(p)$ 接近于 1; 而当 p 接近于 0.5 时, 取 $\omega(p)$ 接近于 0.

三、精度与样本量的关系

1. 用 \bar{x} 估计 $x_{.50}$ 时, 不同精度 d 所需简单随机样本的样本量

估计量的精度与样本量直接有关. 精度要求愈高, 所需的样本量就愈大. 此外, 对精度的不同提法, 计算样本量的方法也不尽相同. 这里我们针对上面第二段中提出的精度定义, 先导出当用 \bar{x} 估计 $x_{.50}$ 时, 简单随机抽样的样本量 n 与给定精度 d 之间的关系.

设 x 的分布遵从 $N(\mu, \sigma^2)$, $x_{0.5} = \mu$, $\bar{x} \sim N(\mu, \sigma^2/n)$. 故

$$p\left\{x_{.50} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \tilde{x}_{.50} \leq x_{.50} + 1.96 \frac{\sigma}{\sqrt{n}}\right\} = 0.95. \quad (11.104)$$

根据(11.101)式对 d 的定义, n 应满足

$$x_{.50} - 1.96 \frac{\sigma}{\sqrt{n}} = x_{.50-d}$$

及

$$x_{.50} + 1.96 \frac{\sigma}{\sqrt{n}} = x_{.50+d}.$$

鉴于正态分布的对称性, 对于给定的 d 值, 上两式确定的 n 相等, 即为

$$n = \left[\frac{1.96}{\frac{x_{.50+d} - x_{.50}}{d}} \right]^2 = \left[\frac{1.96}{u_{.50+d}} \right]^2. \quad (11.105)$$

例如给定 $d=1\%$ 时, $n = \left(\frac{1.96}{0.02507} \right)^2 \doteq 6113$. 对于不同的 d 值, 所需的 n 值如表 11.22 所示.

表 11.22 用 \bar{x} 估计 $x_{.50}$ 时, 不同的精度所需的简单随机抽样的样本量

$d(\%)$	0.5	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8
n	24457	6113	5052	4245	3617	3118	2716	2387	2114	1886

2. 用 \tilde{x}_p 估计 x_p 时, 不同 p 值实际达到的精度

当用简单随机样本的分位数 \tilde{x}_p 估计 x_p 时, 众所周知, \tilde{x}_p 的渐近分布为

$$N\left(x_p, \frac{\sigma^2 p(1-p)}{n\varphi^2(u_p)}\right), \quad (11.106)$$

其中 $\varphi(\cdot)$ 是标准正态分布的密度函数. 下面我们推导对于给定的 n 与 p , 当用 \tilde{x}_p 估计 x_p 时, 实际能达到的精度 d^* .

由(11.106)式, 当 n 大时, 对于 95% 的置信度,

$$\begin{aligned} x_p + \frac{1.96\sigma}{\sqrt{n}} \frac{\sqrt{p(1-p)}}{\varphi(u_p)} &= x_{p+d^*}, \\ \frac{x_{p+d^*} - x_p}{\sigma} &= \frac{1.96}{\sqrt{n}} \frac{\sqrt{p(1-p)}}{\varphi(u_p)}. \end{aligned}$$

从而

$$u_{p+d^*} = u_p + \frac{1.96}{\sqrt{n}} \frac{\sqrt{p(1-p)}}{\varphi(u_p)}. \quad (11.107)$$

或根据(11.105)式, 有

$$u_{p+d^*} = u_p + u_{.50+d} \frac{\sqrt{p(1-p)}}{\varphi(u_p)}. \quad (11.108)$$

由此对给定的 d (或 n) 以及 p , 可计算 d^* 值. 当 d 分别为 1% 与 1.5% 时, 对不同的 p 值, d^* 值如表 11.23 所示.

表 11.23 当 $d=1\%$ 与 $d=1.5\%$ 时不同 p 值 \tilde{x}_p 实际达到的精度 d^* (单位 %)

$p(\%)$	1 (99)	2.5 (97.5)	5 (95)	10 (90)	20 (80)	30 (70)	40 (60)	50
$d=1\%$ 时的 d^*	0.223	0.366	0.523	0.731	0.988	1.139	1.231	1.253
$d=1.5\%$ 时的 d^*	3.317	0.532	0.767	1.032	1.470	1.701	1.831	1.880

在[4]中我们曾列出对不同的 p 值 $k=d^*/d$ 的近似值,实际上,这个值对不同的 d 值并不等于常数,但此值变化很小.例如根据表11.23,我们有 d^*/d ($d=1\%, 1.5\%$)值如表11.24.

表 11.24 当 $d=1\%$ 及 1.5% 时的 d^*/d 值

$p\%$	1 (92)	2.5 (97.5)	5 (95)	10 (90)	20 (80)	30 (70)	40 (60)	50
$d^*/0.01$	0.2228	0.3661	0.5228	0.7314	0.9875	1.1387	1.2350	1.2533
$d^*/0.015$	0.2115	0.3544	0.5116	0.7215	0.9800	1.1337	1.2204	1.2531

再如对 $p=0.50$,及不同的 d , $k=d^*/d$ 的值如表11.25.

表 11.25 $p=0.5$ 时 $k=d^*/d$ 的值

$d(\%)$	1	1.5	2	3	10
d^*/d	1.2533	1.2531	1.2530	1.2514	1.2456

考虑到实际使用,可以将 $k=d^*/d$ 近似当作为仅是 p 的函数.由表11.23知,当取 $d=1\%$ 时,当用 $\hat{x}_{.05}=\tilde{x}_{.05}$ 估计 $x_{.05}$ 时,实际达到的精度为0.523%,也即

$$P\{x_{.04497} < \hat{x}_{.05} < x_{.05523}\} = 95\%.$$

由此可见,当我们用(11.105)式确定简单随机样本量,用(11.103)式估计 x_p 时,对 $p=0.5$, $p \leq 0.2$ 及 $p \geq 0.8$, \hat{x}_p 的实际精度均在 d 之内;而对 $0.2 < p < 0.5$ 或 $0.5 < p < 0.8$,只要适当选取权 $\omega(p)$, \tilde{x}_p 的精度也接近于 d 的水平.

3. 整群抽样的样本量

由(11.105)式确定的样本量 n 仅适用于简单随机抽样或按比例分配的分层随机抽样,对于整群随机抽样,根据(11.96)式,样本量 n' 应为

$$n' = n \cdot \text{deff} = n[1 + (\bar{M} - 1)\rho]. \quad (11.109)$$

例如当要求 $d=1\%$, $\bar{M}=80$, $\rho=0.00775$ 时,

$$n' = 6113 \times [1 + (80 - 1) \times 0.00775] \approx 9856.$$

在实际问题中,考虑到测试记录可能出现的错误以及其他原因,在数据处理时,可能剔除一部分数据,因此我们通常在(11.109)式的基础上增加10%左右的余量.例如在[1]中,方案规定对成年男子及女子两个子总体,分别测量11000人.

四、方案所能达到的绝对精度

在上段中给出了相对精度 d 与样本量 n 或 n' 的关系。本段讨论对于一个已制定的抽样方案(即 d 及 n 或 (n') 已给定), \hat{x}_p 所能达到的绝对精度 Δ

根据(11.99)式, \hat{x}_p 的绝对精度 Δ 满足(在置信度为 0.95 下):

$$P\{\hat{x}_p - x_p \leq \Delta\} = 0.95. \quad (11.110)$$

Δ 的实际值不仅与 n 有关, 而且与 x 的标准差 σ 有关, 当然也与 \hat{x}_p 的估计方法有关。对于中位数 $x_{0.5} = \mu$, 我们是用 x 来估计, 此时

$$\Delta = 1.96 \frac{\sigma}{\sqrt{n}}. \quad (11.111)$$

若用样本分位数 \tilde{x}_p 估计 x_p , 则 Δ 的计算公式为(当 n 大时)

$$\Delta = \frac{1.96\sigma}{\sqrt{n}} \sqrt{\frac{p(1-p)}{\varphi(u_p)}}. \quad (11.112)$$

(11.111)、(11.112)两式中的 σ , 一般用历史资料或试调查资料所获得的估计量代替。

下面我们以某省女性总体为例, 给出几个体形尺寸指标的绝对精度。该方案的相对精度 $d=1.5\%$, 按简单随机抽样的样本量根据表 11.22 为 $n=2716$, 表 11.26 给出了体高、胸围、及腰围的中位数 $x_{0.5}$, $x_{0.2}$ (或 $x_{0.8}$), $x_{0.05}$ (或 $x_{0.95}$) 的估计量所能达到的绝对精度。

表 11.26 某省女性总体抽样方案体高、胸围及腰围分位数估计的绝对精度 Δ

测量指标 x	总体标准差 σ	$\hat{x}_{0.5}$ 的 Δ	$\hat{x}_{0.2}$ 的 Δ	$\hat{x}_{0.05}$ 的 Δ
体高	4.97 cm	0.187 cm	0.267 cm	0.395 cm
胸围	4.90 cm	0.184 cm	0.263 cm	0.389 cm
腰围	6.13 cm	0.231 cm	0.329 cm	0.487 cm

从表 11.26 中的数值可看出, 绝对精度也能满足实际需要, 其他体形指标的标准差一般比这三个指标的标准差都要小, 因此绝对精度更高。

参 考 资 料

- [1] 毕健, 孙山泽, 冯士雍. 中国成年人人体测量抽样方案(技术报告), 1985.
- [2] 孙山泽, 冯士雍, 吴国富. 修订 GB1335-81 服装号型系列标准人体测量抽样方案(技术报告), 1987.
- [3] 冯士雍, 孙山泽, 毕健. 军人体型尺寸测量的抽样方案(技术报告), 1976.
- [4] San Shanze, Feng Shiyong, Bi Jian. Sampling Plans for Human Body Measurements, Sino-American Statistical Meeting, Contributed Papers, Beijing 1987, 403 ~ 405.

评 注

1) 与前面十个案例不同, 本案例讨论的需估计的总体指标是分位数 X_p , 在人体测量抽样时, 这是非常现实的。这是因为人体测量的目的主要是用来制订诸如服装号型标准以及与人体工效有关的各种标准。此时需要估计的是人体各种尺寸的分位数, 而非平均数。

由于估计对象不同, 估计精度的提法也会有所不同。本案例中我们讨论的既非一般意义的绝对误差, 也非通常意义的相对误差, 而是一种与分位数定义有关的另一种意义的相对误差, 即公式(11.101)所定义的 d 。

2) 本例用的抽样方法是分层整群抽样。按人类学标准将全国各省(自治区、直辖市)分为6个自然区域, 并以此为大层。层内以每人的工作(学习)单位作为自然群体(作适当调整使其成为等大小的)进行整群抽样。这也是由于人体测量的实际过程所决定的。不过对群的抽取事实上很难做到严格随机的, 因为难于获得对群的抽样框。整群抽样的优点之一是可用公式比较精确地估计 d_{eff} , 只要知道群内相关 ρ_0 即可。而后者通过试调查, 可用方差分析计算得到。这样一旦计算出简单随机抽样所需的样本量, 即可得到同样精度下整群抽样所需的样本量。

3) 本案例所用的对分位数的估计需要假定总体分布是正态分布的, 或至少要是对称分布的。事实上, 作为一个自然区域内的人群各人体尺寸绝大多数都服从正态分布, 因此文中推荐的估计公式是可用的。不过正如第9章中所指出的那样, 分位数的估计也可用别的方法, 此时就不必对总体分布作特殊的假定, 且有比较好的性质。

附表 随机数表

随机数表(I)

03	47	43	73	86	36	96	47	36	61	46	98	68	71	62	33	26	16	80	45	60	11	14	10	95
97	74	24	67	62	42	81	14	57	20	42	53	32	37	32	27	07	36	07	51	24	51	79	89	73
16	76	62	27	66	56	50	26	71	07	32	90	79	78	53	13	55	38	58	59	88	97	54	14	10
12	56	85	99	26	96	96	68	27	31	05	03	72	93	15	57	12	10	14	21	88	26	49	81	76
50	59	56	35	64	38	54	82	46	22	31	62	43	09	90	06	18	44	32	53	23	83	01	30	30
16	22	77	94	39	49	54	43	54	82	17	37	93	23	78	87	35	20	96	43	84	26	34	91	64
84	42	17	53	31	57	24	55	06	88	77	04	74	47	67	21	76	33	50	25	83	92	12	06	76
62	01	63	78	59	10	95	55	67	19	98	10	50	71	75	12	86	73	58	07	44	39	52	38	79
33	21	12	34	29	78	64	56	07	82	52	42	07	44	38	15	51	00	13	42	99	66	02	79	54
57	60	86	82	44	09	47	27	96	54	49	17	46	09	62	90	52	84	77	27	08	02	73	43	28
18	18	07	92	45	44	17	16	58	09	79	83	86	19	62	06	76	50	03	10	55	23	64	05	05
26	62	38	97	75	84	16	07	44	99	83	11	46	32	24	20	14	85	88	45	10	93	72	88	71
23	42	40	64	74	82	97	77	77	81	07	45	32	14	08	32	98	94	07	72	93	85	79	10	75
52	86	28	19	95	50	92	26	11	97	00	56	76	31	38	80	22	02	53	53	86	60	42	04	53
37	85	94	35	12	83	39	50	08	30	42	34	07	96	88	54	42	06	87	98	35	85	29	48	39
70	29	17	12	13	40	33	20	38	26	13	89	51	03	74	17	76	37	13	04	07	74	21	19	30
56	62	18	37	35	96	83	50	87	75	97	12	25	93	47	70	33	24	03	54	97	77	46	44	80
99	49	57	22	77	88	42	95	45	72	16	64	36	16	00	04	43	18	66	79	94	77	24	21	90
16	08	15	04	72	33	27	14	34	09	45	59	34	68	49	12	72	07	34	45	99	27	72	95	14
31	16	93	32	43	50	27	89	87	19	20	15	37	00	49	52	85	66	60	44	38	68	88	11	80
68	34	30	13	70	55	74	30	77	40	44	22	78	84	26	04	33	46	09	52	68	07	97	06	57
74	57	25	65	76	59	29	97	68	60	71	91	38	67	54	13	58	18	24	76	15	54	55	95	52
27	42	37	86	53	48	55	90	65	72	96	57	69	36	10	96	46	92	42	45	97	60	49	04	91
00	39	68	29	61	66	37	32	20	30	77	84	57	03	29	10	45	65	04	26	11	04	96	67	24
29	94	98	94	24	68	49	69	10	82	53	75	91	93	30	34	25	20	57	27	40	48	73	51	92
16	90	82	66	59	83	62	64	11	12	67	19	00	71	74	60	47	21	29	68	02	02	37	03	31
11	27	94	75	06	06	09	19	74	66	02	94	37	34	02	76	70	90	30	86	38	45	94	30	33
35	24	10	16	20	33	32	51	26	38	79	78	45	04	91	16	92	53	56	16	02	75	50	95	98
38	23	16	86	38	42	38	97	01	50	87	75	66	81	41	40	01	74	91	62	48	51	84	08	32
31	96	25	91	47	96	44	33	49	13	34	86	82	53	91	00	52	43	48	85	27	53	26	89	62
66	67	40	67	14	64	05	71	95	86	11	05	65	09	68	76	83	20	37	90	57	16	00	11	66
14	90	84	45	11	75	73	88	05	90	52	27	41	14	86	22	98	12	22	08	07	52	74	95	80
68	05	51	18	00	33	96	02	75	19	07	60	62	93	55	59	33	82	43	90	49	37	38	44	59
20	46	78	73	90	97	51	40	14	02	04	02	33	31	03	39	54	16	49	36	47	95	93	13	30
64	19	58	97	79	15	06	15	93	20	01	90	10	75	06	40	78	78	89	62	02	67	74	17	33
05	28	93	70	60	22	35	85	15	13	92	03	51	59	77	59	56	78	06	83	52	91	05	70	74
07	97	10	88	23	09	98	42	99	64	61	71	62	99	15	06	51	29	16	93	58	05	77	09	51
68	71	86	85	85	54	87	66	47	54	73	32	08	11	12	44	95	92	63	16	29	56	24	29	48
26	99	61	65	53	58	37	78	80	70	42	10	50	67	42	32	17	55	85	74	94	44	67	16	94
14	65	52	68	75	87	59	36	22	41	26	78	63	06	55	13	08	27	01	50	15	29	39	39	43
17	53	77	58	71	71	41	61	50	72	12	41	94	96	26	44	95	27	36	99	02	96	74	30	83
90	26	59	21	19	23	52	23	33	12	96	93	02	13	39	07	02	18	36	07	25	99	32	70	23
41	23	52	55	99	31	04	49	69	96	10	47	48	45	83	13	41	43	89	20	97	17	14	49	17
60	20	50	81	69	31	99	73	68	68	35	81	33	03	76	24	30	12	48	60	18	99	10	72	34
91	25	38	05	90	94	58	28	41	36	45	37	59	03	09	90	35	57	29	12	82	62	54	65	60
34	50	57	74	37	98	80	33	00	91	09	77	93	19	82	74	94	80	04	04	45	07	31	66	49
85	22	04	39	43	73	81	53	94	79	33	62	46	86	23	08	31	54	46	31	53	94	13	38	47
09	79	13	77	48	73	82	97	22	21	06	03	27	24	83	72	89	44	05	60	35	80	39	94	88
88	75	80	18	14	22	95	75	42	49	39	32	82	22	49	02	48	07	70	37	16	04	61	67	87
90	96	23	70	00	39	00	03	06	90	55	85	78	38	36	94	37	30	69	32	90	89	00	76	33

随机数表(II)

53 74 23 99 67	61 32 28 69 84	94 62 67 86 24	98 33 41 19 95	47 53 53 38 09
63 38 06 86 54	99 00 65 26 94	03 82 90 23 07	79 62 67 80 60	75 91 12 81 19
3 30 58 21 46	05 72 17 19 94	21 21 31 70 96	49 28 14 00 49	55 65 79 78 07
63 43 36 82 62	65 51 18 37 88	61 38 44 12 45	32 92 83 88 60	54 34 81 85 35
98 25 37 55 26	01 91 82 81 46	74 71 12 94 97	24 02 71 37 07	03 92 18 66 75
02 63 21 17 62	71 50 80 89 56	38 15 70 11 48	43 40 45 86 95	00 80 26 91 03
64 55 22 21 82	48 22 28 06 00	61 54 13 43 91	82 78 12 23 29	06 66 24 12 27
85 07 26 13 89	01 10 07 82 04	59 63 69 96 03	69 11 15 83 80	13 29 54 19 28
18 54 16 24 15	51 64 44 82 00	62 61 65 04 69	38 18 65 18 97	85 72 13 49 21
34 85 27 84 87	61 48 64 56 26	90 18 48 18 26	37 70 15 42 57	65 65 80 39 07
63 92 18 27 46	57 99 16 96 56	30 33 72 85 22	84 64 38 56 98	99 01 30 98 64
62 53 30 27 59	37 75 41 66 48	85 97 80 61 45	23 53 04 01 63	45 76 08 64 27
08 45 93 15 22	60 21 75 46 91	93 77 27 83 42	28 88 61 08 84	69 62 03 42 73
07 08 51 18 40	45 44 75 13 90	24 94 96 61 02	57 55 66 83 15	73 42 37 11 61
01 85 89 95 66	51 10 19 34 88	15 84 97 13 75	12 76 39 43 78	64 63 91 08 25
72 84 71 14 35	19 11 58 49 26	50 11 17 17 76	86 31 57 20 18	95 60 78 46 75
88 78 28 16 84	13 52 53 94 53	75 45 69 30 96	73 89 65 70 31	99 17 43 48 76
45 17 75 65 57	28 40 19 72 12	25 12 74 75 67	60 40 60 81 19	24 62 01 61 16
96 76 28 12 54	22 01 11 94 25	71 96 16 16 88	68 64 36 74 45	19 59 50 88 92
43 31 67 72 30	24 02 94 08 63	38 32 36 66 02	69 36 38 25 39	48 03 45 13 22
50 44 66 44 21	66 06 58 05 62	68 15 54 36 02	42 35 48 96 32	14 52 41 52 43
22 66 22 15 86	26 63 75 41 99	53 42 36 72 24	58 37 52 18 51	03 37 18 39 11
16 24 40 14 51	23 22 30 88 57	95 67 47 29 83	94 69 40 06 07	18 16 36 78 86
31 73 91 61 19	60 20 72 93 48	98 57 07 23 69	65 95 39 69 58	56 80 30 19 44
78 60 73 99 84	43 89 94 36 45	56 69 47 07 41	90 22 91 07 12	78 33 54 08 72
84 37 90 61 56	70 10 23 98 05	85 11 34 76 60	76 48 45 34 60	01 64 18 39 36
36 67 10 08 23	98 93 35 08 86	96 29 76 29 81	33 34 91 58 93	63 14 52 32 52
07 28 59 07 48	89 64 58 89 75	83 85 62 27 89	30 14 78 56 27	86 63 59 80 02
10 15 83 87 60	79 24 31 66 56	21 48 24 06 93	91 98 94 05 49	01 47 59 38 00
50 19 68 97 63	03 73 52 16 56	00 53 55 90 27	33 42 29 38 87	22 13 88 83 34
53 81 29 13 39	35 01 20 71 34	62 33 74 82 14	53 73 19 09 03	56 54 29 56 23
51 80 32 08 92	33 93 74 66 99	40 14 71 94 58	45 94 19 38 81	14 44 99 81 07
31 91 70 29 13	80 03 54 07 27	96 94 78 32 66	50 95 52 74 33	13 80 50 62 34
37 71 67 95 13	20 02 44 95 94	64 85 04 05 72	01 32 90 76 14	53 89 74 60 41
93 66 13 83 27	92 79 64 64 72	28 54 96 53 84	48 14 52 98 94	56 07 93 89 30
02 96 08 45 65	13 05 00 41 84	93 07 54 72 59	21 45 37 09 77	19 48 56 27 44
49 83 43 48 35	82 88 38 69 96	72 36 64 19 76	47 45 13 18 60	82 11 38 95 97
84 60 71 62 46	40 80 81 30 37	34 39 23 05 38	25 15 35 71 90	88 12 57 21 77
18 17 30 88 71	44 91 14 88 47	89 23 30 63 15	56 34 20 47 89	99 82 93 24 93
79 69 10 61 78	71 32 76 95 62	87 00 22 58 40	92 54 01 75 21	43 11 71 99 31
75 93 36 57 83	56 20 14 82 11	74 21 97 90 65	96 42 63 63 86	74 54 13 26 94
38 30 32 29 03	06 28 81 39 38	62 25 06 84 63	61 29 08 93 67	04 32 92 03 09
51 29 50 10 34	31 57 75 95 80	51 97 02 74 77	76 13 48 49 44	18 55 63 77 09
21 31 38 86 24	37 79 81 53 74	73 24 16 10 33	52 83 90 94 76	70 47 14 54 36
29 01 23 87 88	58 02 39 37 67	42 10 14 20 92	16 55 23 42 45	54 96 09 11 06
91 38 95 22 00	18 74 72 00 18	38 79 58 69 32	81 76 80 26 92	82 80 84 25 39
90 84 60 79 80	24 36 59 87 38	82 07 13 89 35	96 35 23 79 18	05 98 90 07 35
46 40 62 98 82	54 97 26 56 95	15 74 80 03 32	16 46 70 50 80	67 72 16 42 79
20 31 89 03 43	38 46 82 60 72	32 14 82 99 70	80 60 47 18 97	63 49 30 21 30
71 59 73 05 56	05 12 23 71 77	9 0 13 20 49	82 57 59 26 91	63 29 67 98 60

随机数表(III)

22 17 68 65 84	68 95 23 92 35	87 02 22 57 51	61 09 43 95 06	58 24 82 03 47
15 36 27 59 46	13 79 93 37 55	39 77 32 77 09	85 52 05 30 62	47 83 51 62 74
16 77 23 02 77	09 61 87 25 21	28 06 24 25 93	16 71 13 59 78	23 05 47 47 25
78 43 76 71 61	20 44 90 32 64	97 67 63 99 61	46 38 03 98 22	69 81 21 99 21
03 28 28 26 08	78 37 32 04 05	69 30 10 09 05	88 69 58 28 99	35 07 44 75 47
93 22 53 64 30	07 10 63 76 35	87 03 04 79 88	08 13 13 85 51	55 34 57 72 69
78 76 58 54 74	92 38 70 96 92	52 06 79 79 45	82 68 18 27 44	69 66 92 19 09
23 68 35 26 00	99 53 93 61 28	52 70 05 48 34	56 65 05 61 86	90 92 10 70 80
15 39 25 70 99	93 86 52 77 65	15 33 59 05 28	22 87 26 07 47	86 96 98 29 06
58 71 96 30 24	18 46 23 34 27	85 13 99 24 44	49 18 09 79 49	74 16 32 23 02
57 35 27 33 72	24 53 63 94 09	41 10 76 47 91	44 04 95 49 66	39 60 04 59 81
48 50 86 54 48	22 06 34 72 52	82 21 15 65 20	33 29 94 71 11	15 91 29 12 03
61 96 48 95 03	07 16 39 33 66	98 56 10 56 79	77 21 30 27 12	90 49 22 23 62
36 93 89 41 26	29 70 83 63 51	99 74 20 52 36	87 09 41 15 09	98 60 16 63 03
18 87 00 42 31	57 90 12 02 07	23 47 37 17 31	54 08 01 88 63	39 41 88 92 10
88 56 53 27 59	33 35 72 67 47	77 34 55 45 70	08 18 27 38 90	16 95 86 70 75
09 72 95 84 29	49 41 31 06 70	42 38 06 45 18	64 84 73 31 65	52 53 37 97 15
12 96 88 17 31	65 19 69 02 83	60 75 86 90 68	24 64 19 35 51	56 61 87 39 12
85 94 57 24 16	92 09 84 38 76	22 00 27 69 85	29 81 94 78 70	21 94 47 90 12
38 64 43 59 98	98 77 87 68 07	91 51 67 62 44	40 98 05 93 78	23 32 65 41 18
53 44 09 42 72	00 41 86 79 79	68 47 22 00 20	35 55 31 51 51	00 83 63 22 50
40 76 66 26 84	57 99 99 90 37	36 63 32 08 58	37 40 13 68 97	87 64 81 07 83
02 17 79 18 05	12 59 52 57 02	22 07 90 47 03	28 14 11 30 79	20 69 22 40 98
9 17 82 06 53	31 51 10 96 46	92 06 88 07 77	56 11 50 81 69	40 23 72 51 39
35 76 22 42 92	96 11 83 44 80	34 68 35 48 77	33 42 40 90 60	73 96 53 97 86
26 29 13 56 41	85 47 04 66 08	34 72 57 59 13	82 43 80 46 15	38 26 61 70 04
77 80 20 75 82	72 82 32 99 90	63 95 73 76 63	89 73 44 99 05	48 67 26 43 18
46 40 56 44 52	91 36 74 43 53	30 82 13 54 00	78 45 63 98 35	55 03 36 67 68
37 56 08 18 09	77 53 84 46 47	31 91 18 95 58	24 16 74 11 53	44 10 13 85 57
61 65 61 63 66	37 27 47 39 19	84 83 70 07 48	53 21 40 06 71	95 06 79 88 54
93 43 69 64 07	34 13 04 52 35	56 27 09 24 86	61 85 53 83 45	19 90 70 99 00
21 96 60 12 99	11 20 99 45 18	48 13 93 55 34	18 37 79 49 60	65 97 38 20 46
95 20 47 97 97	27 37 83 28 71	00 06 41 41 74	45 89 09 39 84	51 67 11 52 49
97 86 21 78 73	10 65 81 92 59	58 76 17 14 97	04 76 62 16 17	17 95 70 45 80
69 92 06 34 13	59 71 74 17 32	27 55 10 24 19	28 71 82 13 74	63 52 52 01 41
04 31 17 21 56	33 73 99 19 87	26 72 39 27 67	53 77 57 68 93	60 61 97 22 61
61 06 98 03 91	87 14 77 43 96	43 00 65 98 50	45 60 33 01 07	98 99 46 50 47
85 93 85 86 88	72 87 08 62 40	16 06 10 89 20	23 21 34 74 97	76 38 03 29 63
21 74 32 47 45	73 96 07 94 52	09 65 90 77 47	25 76 16 19 33	53 05 70 53 30
15 69 53 82 80	79 96 23 53 10	65 39 07 16 29	45 33 02 43 70	02 87 40 41 45
32 89 08 04 49	20 21 14 68 86	87 63 93 95 17	11 29 01 95 80	35 14 97 35 33
87 18 15 89 79	85 43 01 72 73	08 61 74 51 69	89 74 39 82 15	94 51 33 41 67
98 83 71 94 22	59 97 50 99 52	08 52 85 08 40	87 80 61 65 31	91 51 80 32 44
10 03 58 21 66	72 68 49 29 31	89 85 84 46 06	59 73 19 85 23	65 09 29 75 63
47 90 56 10 08	88 02 84 27 83	42 29 72 23 19	66 56 45 65 79	20 71 53 20 25
22 85 61 68 90	49 64 92 85 44	16 40 12 89 88	50 14 49 81 06	01 82 77 45 12
67 80 43 79 33	12 83 11 41 16	25 58 19 68 70	77 02 54 00 52	53 43 37 15 26
27 62 50 96 72	79 44 61 40 15	14 53 40 65 39	27 31 58 50 28	11 39 03 25 25
33 78 80 87 15	38 30 06 38 21	14 47 47 07 26	54 96 87 53 32	40 36 40 96 76
13 13 92 66 99	47 24 49 57 74	32 25 43 62 17	10 97 11 69 84	99 63 22 32 93

随机数表(IV)

10 27 53 96 23	71 50 54 36 23	54 31 04 82 98	04 14 12 15 09	26 78 25 47 47
28 47 50 61 88	64 85 27 20 18	83 36 36 05 56	39 71 65 09 62	94 76 62 11 89
34 21 42 57 02	59 19 13 97 48	80 30 03 30 98	05 24 67 70 07	84 97 50 87 46
61 81 77 23 23	82 82 11 54 08	53 28 70 58 96	44 07 39 55 43	42 34 43 39 28
61 15 18 13 54	16 86 20 26 88	90 74 80 55 09	14 53 90 51 17	52 01 63 01 59
91 76 21 64 64	44 91 13 32 97	75 31 62 66 54	84 80 92 75 77	56 08 25 70 29
00 97 79 08 06	37 30 28 59 85	53 56 68 53 40	01 74 39 59 73	30 19 99 85 48
36 46 18 34 94	75 20 50 27 77	78 91 69 16 00	08 43 18 73 68	67 69 61 34 25
88 98 99 60 50	65 95 79 42 94	93 62 40 89 96	43 56 47 71 66	46 76 29 67 02
04 37 59 87 21	05 02 03 24 17	47 97 81 56 51	92 34 86 01 82	55 51 33 12 91
63 62 06 34 41	94 21 78 55 09	72 76 45 16 94	29 95 81 83 83	79 88 01 97 30
78 47 23 53 90	34 41 92 45 71	09 23 70 70 07	12 38 92 79 43	14 85 11 47 20
87 68 62 15 43	53 14 36 59 25	54 47 33 70 15	59 24 48 40 35	50 03 42 99 36
47 60 92 10 77	83 59 53 11 52	66 25 69 07 04	48 68 64 71 06	61 65 70 22 12
56 88 87 59 41	65 28 04 67 53	95 79 88 37 31	50 41 06 94 76	81 83 17 16 33
02 57 45 86 67	73 43 07 34 48	44 26 87 93 29	77 09 61 67 84	06 69 44 77 75
31 54 14 13 17	48 62 11 90 60	68 12 93 64 28	46 24 79 16 76	14 60 25 51 01
28 50 16 43 36	28 97 85 58 99	67 22 52 76 23	24 70 36 54 54	59 28 61 71 96
63 29 62 66 50	02 63 45 52 38	67 63 47 54 75	83 24 78 43 20	92 63 13 47 48
45 65 58 26 51	76 96 59 38 72	86 57 45 71 46	44 67 76 14 55	44 88 01 62 12
39 65 36 63 70	77 45 85 50 51	74 13 39 35 22	30 53 36 02 95	49 34 88 73 61
73 71 98 16 04	29 18 94 51 23	76 51 94 84 86	79 93 96 38 63	08 58 25 38 94
72 20 56 20 11	72 65 71 08 86	79 57 95 13 91	97 48 72 66 48	09 71 17 24 89
75 17 26 99 76	89 37 20 70 01	77 31 61 95 46	26 97 05 73 61	53 33 18 72 87
7 48 60 82 29	81 30 15 39 14	48 38 75 93 29	06 87 37 78 48	45 56 00 84 47
68 08 02 30 72	83 71 43 30 49	89 17 95 88 29	02 39 56 03 46	97 74 06 56 17
14 23 98 61 67	70 52 85 01 50	01 34 02 78 43	10 62 98 19 41	18 83 99 47 99
49 03 96 21 44	25 27 99 41 28	07 41 08 34 66	19 42 74 39 91	41 96 53 78 72
78 37 06 08 43	63 61 62 42 29	39 68 95 10 96	09 24 23 00 62	56 12 30 73 16
37 21 34 17 68	68 96 83 23 56	32 84 60 15 31	44 73 67 34 77	91 15 79 74 58
14 29 09 34 04	87 83 07 55 07	76 58 30 83 64	87 29 25 58 84	86 50 60 00 25
58 43 28 06 36	49 52 83 51 14	47 56 91 29 34	05 87 31 06 95	12 45 57 09 09
10 43 67 29 70	80 62 80 03 42	10 80 21 38 84	90 56 35 03 09	43 12 74 49 14
44 38 38 39 54	86 97 37 44 22	00 95 01 31 76	17 16 29 56 63	38 78 94 49 81
90 69 59 19 51	85 39 52 85 13	07 28 37 07 61	11 16 36 27 03	78 86 72 04 95
41 47 10 25 62	97 05 31 03 61	20 26 36 31 62	68 69 86 95 44	84 95 48 46 45
91 94 14 63 19	75 89 11 47 11	31 56 34 19 09	79 57 92 36 59	14 93 87 81 40
80 06 54 18 66	09 18 94 06 19	98 40 07 17 81	22 45 44 84 11	24 62 20 42 31
67 72 77 63 48	84 08 31 55 58	24 33 45 77 58	80 45 67 93 82	75 70 16 03 24
59 40 24 13 27	79 26 88 86 30	01 31 60 10 39	53 58 47 70 93	85 81 56 39 38
05 90 35 89 95	01 61 16 96 94	50 78 13 69 36	37 68 53 37 31	71 26 35 03 71
44 43 80 69 98	46 68 05 14 82	90 78 50 05 62	77 79 13 57 44	59 60 10 39 66
61 81 31 96 82	00 57 25 60 59	46 72 60 18 77	55 66 12 62 11	08 99 55 64 57
42 88 07 10 05	24 98 65 63 21	47 21 61 88 32	27 80 30 21 60	10 92 35 36 12
77 94 30 05 39	28 10 99 00 27	12 79 73 99 12	49 99 57 94 82	96 88 57 17 91
78 88 19 76 16	94 11 68 84 26	23 54 20 86 85	23 86 55 99 07	36 37 34 92 09
87 76 59 61 81	43 63 64 61 61	65 76 36 95 90	18 48 27 45 68	27 23 65 30 72
91 43 05 96 47	55 78 99 95 24	37 55 85 78 78	01 48 41 19 10	35 19 54 07 73
84 97 77 72 73	09 62 06 65 72	87 12 49 03 60	41 15 20 76 27	50 47 02 29 16
87 41 60 76 83	44 88 96 07 80	83 05 83 38 96	73 70 86 81 90	30 56 10 48 59

随机数表(V)

28 89 65 87 08	13 50 63 04 23	25 47 57 81 13	52 62 24 19 94	91 67 48 57 10
30 29 43 65 42	78 66 28 55 80	47 46 41 90 08	55 98 78 10 70	49 92 05 12 07
95 74 62 60 53	51 57 32 22 27	12 72 72 27 77	44 67 32 23 13	67 95 07 76 30
01 85 54 96 72	66 86 65 64 60	56 59 75 36 75	46 44 33 63 71	54 50 06 44 70
10 91 46 96 86	19 83 52 47 53	65 00 51 93 51	30 80 05 19 29	56 23 27 19 03
05 33 18 08 51	51 78 57 26 17	34 87 96 23 95	89 99 93 39 79	11 28 94 15 52
04 43 13 37 00	79 68 96 26 60	70 39 83 66 56	62 03 55 86 57	77 55 33 62 03
05 85 40 25 24	73 52 93 70 50	48 21 47 74 63	17 27 27 51 26	35 96 29 00 45
84 90 90 65 77	63 99 25 69 02	09 04 03 35 78	19 79 95 07 21	02 84 48 51 97
28 55 53 09 48	86 28 30 02 35	71 30 32 06 47	93 74 21 86 33	49 90 21 69 74
89 83 40 69 80	97 96 47 59 97	56 33 24 87 36	17 18 16 90 46	75 27 28 52 13
73 20 96 05 68	93 41 69 96 07	97 50 81 79 59	42 37 13 81 83	92 42 85 04 31
10 89 07 76 21	40 24 74 36 42	40 33 04 46 24	35 63 02 31 61	34 59 43 36 96
91 50 27 78 87	06 06 16 25 98	17 78 80 36 85	26 41 77 63 37	71 63 94 94 33
03 45 44 66 88	97 81 26 03 89	39 46 67 21 17	98 10 89 33 15	61 63 00 23 92
89 41 58 91 63	65 99 59 97 84	90 14 79 61 55	56 16 88 87 60	32 15 99 67 43
13 43 00 97 26	16 91 21 32 41	60 22 66 72 17	31 85 33 69 07	68 49 20 43 29
71 71 00 51 72	62 03 89 26 32	35 27 99 18 25	78 12 03 09 70	50 93 19 35 56
19 28 15 00 41	92 27 73 40 38	37 11 05 75 16	98 81 99 37 29	92 20 32 39 67
56 38 30 92 30	45 51 94 69 04	00 84 14 36 37	95 66 39 01 09	21 68 40 95 79
39 27 52 89 11	00 81 06 28 48	12 08 05 75 26	03 35 63 05 77	13 81 20 67 53
73 13 28 58 01	05 06 42 24 07	60 60 29 99 93	72 93 78 04 36	25 76 01 54 03
81 60 84 51 57	12 68 46 55 89	60 09 71 87 89	70 81 10 95 91	83 79 68 20 66
05 62 98 07 85	07 79 26 69 61	67 85 72 97 41	85 79 76 48 23	61 58 87 08 05
62 97 16 29 18	52 16 16 23 56	62 95 80 97 63	32 25 34 03 36	48 84 60 37 65
31 13 63 21 08	16 01 92 58 21	48 79 74 73 72	08 64 80 91 38	07 28 66 61 59
97 38 35 34 19	89 84 05 34 47	88 09 31 54 88	97 96 86 01 69	46 13 95 65 96
32 11 78 33 82	51 99 98 44 39	12 75 10 60 36	80 66 39 94 97	42 36 31 16 59
81 99 13 37 05	03 12 60 39 23	61 73 84 89 18	26 02 04 37 95	96 18 69 06 30
45 74 00 03 05	69 99 47 26 52	48 06 30 00 18	03 30 28 55 69	66 10 71 44 05
11 84 13 69 01	88 91 28 79 50	71 42 14 96 55	98 59 96 01 36	88 77 90 45 59
14 66 12 87 23	59 45 27 03 51	85 64 23 85 41	64 72 08 59 44	67 98 36 65 56
40 25 67 87 82	84 27 17 30 37	48 69 49 02 58	98 02 50 58 11	95 39 06 35 63
44 48 97 49 43	65 45 53 41 07	14 83 46 74 11	76 66 63 60 08	90 54 33 65 84
41 94 54 06 57	48 28 01 83 84	09 11 21 91 73	97 28 44 74 06	22 30 95 69 72
07 12 35 58 84	93 18 31 83 45	54 52 62 29 91	53 58 54 66 05	47 19 63 92 75
64 27 90 43 52	18 26 32 96 83	50 58 45 27 57	14 96 39 64 85	73 87 96 76 23
80 71 86 41 03	45 62 63 40 88	35 69 34 10 94	32 22 52 04 74	69 68 21 83 41
27 06 08 09 92	26 22 59 28 27	38 58 22 14 79	24 32 12 38 42	33 56 90 92 57
54 68 97 20 54	33 26 74 03 30	74 22 19 13 43	30 28 01 92 49	58 61 52 27 03
02 92 65 68 99	05 53 15 26 70	04 69 22 64 07	04 73 25 74 82	78 35 22 21 88
83 52 57 78 62	98 61 70 48 22	68 50 64 55 75	42 70 32 09 60	58 70 61 43 97
82 82 76 31 33	85 13 41 38 10	16 47 61 43 77	83 27 19 70 41	34 78 77 60 25
38 61 34 09 49	04 41 66 09 76	20 50 73 40 95	24 77 95 73 20	47 42 80 61 03
01 01 11 88 38	03 10 16 82 24	39 58 20 12 39	82 77 02 18 88	33 11 49 15 16
21 66 14 38 28	54 08 18 07 04	92 17 63 36 75	33 14 11 11 78	97 30 53 62 00
32 29 30 69 59	68 50 33 31 47	15 64 88 75 27	04 51 41 61 96	86 62 93 66 71
04 59 21 65 47	39 90 89 36 77	46 86 86 38 86	50 09 13 24 91	54 80 67 78 00
38 64 50 07 36	56 50 45 94 25	48 28 48 30 51	60 73 73 03 87	68 47 37 10 84
48 33 50 83 53	39 77 64 59 90	58 92 62 50 18	93 09 45 89 06	13 26 98 86 29

参 考 文 献

- [1] 中华人民共和国卫生部. 国家卫生服务总调查方案及调查指导手册, 1993
- [2] 中华人民共和国国家标准 GB10111 88, 利用随机数骰子进行随机抽样的方法, 中国标准出版社, 1989
- [3] 冯士雍, 抽样调查的设计与分析 I, II, 数理统计与管理, 1993, 12(1): 46~51, 12(2): 53~57
- [4] 冯士雍, 王思平. 1987 年中国儿童情况抽样调查的抽样设计及数据处理模式. 中国儿童状况的调查与研究. 中国统计出版社, 1990, 32~46
- [5] 冯士雍, 孙山泽, 毕健. 人体测量抽样方案目标量的估计及样本量的确定. 应用概率统计, 1989, 5: 350~357
- [6] 冯士雍, 杨若勇. 北京地区专业技术人员现状抽样调查的抽样设计, 数据处理方法和精度分析. 应用概率统计, 1991, 7: 425~432
- [7] 冯士雍, 程翰生. 一种科学的统计调查方法——抽样调查. 中国科技论坛, 1986, 第二期: 34~40
- [8] 冯士雍, 程翰生, 汪仁官. 全国粮食污染调查抽样方案的设计与数据处理方法. 应用概率统计, 1989, 1: 155~160
- [9] 许宝騄. 抽样论. 北京大学出版社, 1982
- [10] 林少宫. 抽样调查. 中国大百科全书, 数学, 1988, 84~86
- [11] 国家统计局. 1993 年人口变动情况抽样调查方案, 1993
- [12] 国家统计局农村社会经济调查总队. 农村抽样调查基础工作规程. 北京科学技术出版社, 1990
- [13] 施锡钰. 非参数统计中的刀切法(Jackknife). 应用概率统计, 1987, 69~76
- [14] 莫里斯 H, 汉森, 托尔·戴伦纽斯, 本杰明 J, 特平著; 龚盛尧译. 调查抽样的回顾与前瞻.
- [15] 高嘉陵, 冯士雍. 中国 1986 年 74 城镇人口迁移抽样调查目标量估计与精度分析. 中国人口科学, 1991, 第二期: 1~8
- [16] 陶春芳, 蒋永萍主编. 中国妇女社会地位概观. 中国妇女出版社, 1993
- [17] Beale E M L. Some uses of computers in operational research. Industrielle Organisation, 1962, 31: 51~52
- [18] Beardwood J, Halton J H, Hammersley J M. The shortest path through many points, Proc. Cambridge Phil. Soc., 1959, 55: 299~327
- [19] Bellhouse D R, Rao J N K. Systematic sampling in the presence of a trend. Biometrika, 1975, 62: 694~697
- [20] Bhargava R C. A property of the jackknife estimation of the variance when more than one observation is omitted. Sankhya Ser. A, 1983, 45: 112~119
- [21] Booth G, Sarneck J. Planning some two-factor comparative surveys. Jour Amer Stat Assoc, 1969, 64: 560~573
- [22] Brewer K W R. Ratio estimation in finite populations: Some results deducible from the assumption of an underlying stochastic process. Australina Jour Stat, 1963, 5: 93~105
- [23] Brewer K W R. A sample procedure for sampling π power. Austral J Statist, 1975, 17(3): 166~172

- [24] Brewer K W R, Early L J, Joyce S F. Selecting several samples from a single population. *Austral J Statist*, 1972, 14: 231~239
- [25] Brewer K W R, Hanif M. Sampling with Unequal Probabilities. *Lecture Note in Statistics*. Springer-Verlag, 1983
- [26] Cameron J M. Use of variance components in preparing schedules for the sampling of baled wool. *Biometrics*, 1951, 7: 83~96
- [27] Cassel C M, Sarndal C E, Wretman J H. *Foundations of Inference in Survey Sampling*, John Wiley & Sons, 1977
- [28] Chatterjee S. A note on optimum stratification. *Skand Akt*, 1967, 50: 40~44
- [29] Cochran W G. Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann Math Stat.*, 1946, 16: 164~177
- [30] Cochran W G. *Sampling Techniques*, 3rd ed. John Wiley & Sons, 1977
- [31] Cornfield J. On samples from finite populations. *Jour Amer Stat Assoc*, 1944, 39: 236~239
- [32] Cox D R. Estimation by double sampling. *Biometrika*, 1952, 39: 217~227
- [33] Dalenius T, Hodges J L Jr. Minimum variance stratification. *Jour Amer Stat Assoc*, 1959, 54: 88~101
- [34] Daniels H E. Tail probability approximations. *Int Stat Rev*, 1987, 54: 34~48
- [35] Das A C. On two phase sampling and sampling with varying probabilities. *Full Internal Statist Inst*, 33, Book2, 1951, 105~112
- [36] David F N, Neyman J. Extension of the Markoff theorem of least squares. *Stat. Res. Mem.* 1938, 2: 105
- [37] David I P, Sukhatme B V. On the bias and mean square error of the ratio estimator. *Jour Amer Stat Assoc*, 1974, 69: 464~466
- [38] Davison A C, Hinkley D V. Saddlepoint ap, roximations in resampling methods. *Biometrika*, 1988, 75: 417~431
- [39] Deming W E. On a probability mechanism to attain an economic balance between the resultant error of non-response and the bias non-response. *J A S A*, 1953, 48, 743~772
- [40] Des Raj. Some estimators in sampling with varying probabilities without replacement. *Jour Amer Stat Assoc*, 1956, 51: 269~284
- [41] Des Raj. Some remarks on a simple procedure of sampling without replacement. *Jour Amer Stat Assoc*, 1966, 61: 391~396
- [42] Durbin J. Some results in sampling theory when the units are selected with unequal probabilities. *Jour. Roy. Stat. Soc.*, 1953, B15: 262~269
- [43] Durbin J. A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 1959, 46: 477~480
- [44] Durbin J. Design of multi-stage surveys for the estimation of sampling errors. *App Stat*, 1967, 16: 152~164
- [45] Efron B. Better bootstrap confidence intervals. *Jour Amer Stat Assoc*, 1987, 82: 171~185
- [46] Efron B, Stein C. The Jackknife estimate of variance. *Ann Stat* 1981, 9: 586~596
- [47] Fellegi T. Sampling with varying probabilities without replacement: rotating and non rotating samples. *Jour Amer Stat Assoc*, 1963, 58: 183~201

- [48] Fisher R A, Yates F. Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, Edinburgh, fifth edition, 1957
- [49] Fleiss J L. Statistical Methods for Rates and Proportions, 2nd ed. New York, John Wiley & Sons, 1981
- [50] Fuller W A. Regression analysis for sample surveys. *Sankhya* c, 1975, 37: 117~132
- [51] Godambe V P. A unified theory of sampling from finite populations. *Jour. Roy. Stat Soc*, 1955, B17: 269~278
- [52] Gray H L, Schucany W R. The Generalized Jackknife Statistic. New York: Marcel Dekker, 1972.
- [53] Hajek J. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann Math Statist*. 1964, 35: 1491~1523
- [54] Hajek J. Sampling from a finite population, Marcel Dekker, Inc. 1981
- [55] Haldane J B S. On a method of estimating frequencies. *Biometrika*, 1945, 33: 222~225
- [56] Hansen M H, Hurwitz W N. One the theory of sampling from finite populations. *Ann Math Stat*, 1943, 14: 333~362
- [57] Hansen M H, Hurwitz W N, Madow W G. Sample Survey Methods and Theory. New York, John Wiley and Sons, Vols. I and II. 1953
- [58] Hanurav T V. Optimum utilization of auxiliary information: π PS sampling of two units from a stratum. *Jour. Roy Statist Soc Ser*, 1967, B29: 374~391
- [59] Hanurav T V. Optimum utilization of auxiliary information, π PS sampling of two units from a stratum. (Addenda and corrigenda). *Jour Roy Statist Soc Ser.*, 1969, B31: 192~194
- [60] Hartley H O, Rao J N K, Sampling with unequal probabilities and without replacement. *Ann Math Stat*, 1962, 33: 350~374
- [61] Hartley H O, Ross A. Unbiased ratio estimates. *Nature*, 1954, 174: 270~271
- [62] Holt D, Smith T M F. Post stratification, *Jour Roy Statist Soc*, 1979, A142: 33~46
- [63] Horvitz D G, Shah B V and Simmons W R. The unrelated randomized response model. *Proc Soc stat Sect Amer. Stat. Assoc*. 1967, 65~72
- [64] Horvitz D G, Thompson D J. A generalization of sampling without replacement from a finite universe. *Jour Amer Stat Assoc*, 1952, 47: 663~685
- [65] Jolliffe F R. Survey design and analysis Ellis Horwood Limited. 1986
- [66] Kalton G. Introduction to Survey Sampling. Beverly Hills, Sage Publications. 1983
- [67] Kish L. Survey Sampling. New York. John Wiley & Sons, 1965
- [68] Konijn H S. Statistical Theory of Sample Survey Design and Analysis. London, North Holland. 1973
- [69] Lahiri D B. A method for sample selection providing unbiased ratio estimates. *Bull Int Stat Inst.*, 1951, 33 (2): 133~140
- [70] McCarthy P J. Replication: An Approach to the analysis of data from complex surveys. National Center for Health Statistics, Washington D C, Series, 1966. 2 (14)
- [71] Mickey M R. Some finite population unbiased ratio and regression estimators.

- Jour Amer Stat Assoc., 1959, 54: 594~612
- [72] Midzuno H. On the sampling system with probability proportionate to sum of sizes. Ann Inst Stat Math, 1951, 2, 99~108
- [73] Miller R G, Jr. The Jackknife: A Review Biometrika. 1974, 61: 1~15
- [74] Murthy M N. Ordered and unordered estimators in sampling without replacement Sankhya, 1957, 18: 379~390
- [75] Murthy M N. Sampling Theory and Methods. Statistical Publishing Society, Calcutta, India. 1967
- [76] Narain R D. On sampling without replacement with varying probabilities. Jour. Ind Soc Agric Stat, 1951, 3: 169~174
- [77] Neyman J. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Jour Roy Stat Soc, 1934, 97: 558~606
- [78] Neyman J. Contribution to the theory of sampling human populations. Jour. Amer Stat Assoc, 1938, 33: 101~116
- [79] Ogas J L, Clark D F. The annual survey of manufacturers: a report on methodology. Technical Paper No. 24, US Bureau of the Census, 1971
- [80] Olkin I. Multivariate ratio estimation for finite populations. Biometrika, 1958, 45: 154~165
- [81] Plackett R L, Burman J P. The design of optimum multifactorial experiments, Biometrika 1946, 33: 305~325
- [82] Politz A N, Simmow W R. An attempt to get the "not at homes" into the Sample without callbacks. Jour Amer Stat Assoc. 1949, 1950, 44, 91 and 45: 136~137
- [83] Quenouille M H. Problems in plane sampling. Ann Math Stat, 1949, 20: 355~375
- [84] Quenouille M H. Notes on bias in estimation. Biometrika, 1956, 43: 353~360
- [85] Rao J N K. On the estimation of the relative efficiency of sampling procedures, Ann Inst Stat Math, 1952, 14: 143~150
- [86] Rao J N K. On two simple schemes of unequal probability sampling without replacement, Jour Ind Stat Assoc, 1965, 3: 173~180
- [87] Rao J N K. Ratio and regression estimators. New Developments in Survey Sampling, N. L. Johnson and H. Smith, Jr. (eds.) New York, John Wiley & Sons, 1969, 213~234
- [88] Rao J N K. On double sampling for stratification and analytical surveys. Biometrika, 1973, 60: 125~133
- [89] Rao J N K. Unbiased variance estimation for multistage designs. Sankhya, 1975
- [90] Rao J N K, Hartley H O, Cochran W G. A simple procedure of unequal probability sampling without replacement. Jour Roy Stat Soc, 1962, B24: 482~491
- [91] Reid N. Saddlepoint methods and statistical inference. Stat: Scis, 1988, 3: 213~238
- [92] Sampford M R. On sampling without replacement with unequal probabilities of selection. Biometrika, 1967 54: 499~513
- [93] Sarndal C E. Sample survey theory vs. general statistical theory: Estimation of the population mean, Rev Int Stat Inst, 1972, 40: 1~12
- [94] Scott A, Wu C F. On the asymptotic distribution of ratio and regression

- estimators, Jour Amer Ass. 1981, 76: 98~102
- [95] Sen A R. On the estimate of variance in sampling with varying probabilities. Jour Ind Soc Agric Stat, 1953, 5: 119~127
- [96] Sethi V K. On optimum pairing of units. Sankhya, 1965, B27: 315~320
- [97] Shao J, Shi X Q. Half-Sample variance estimation. Commun Stat Theory Meth. 1989, 18: 4197~4210
- [98] Shi X Q, Wu C F J, Chen J H. Weak and Strong representations for quantile processes from finite populations with application to simulation size in resampling inference. Canadian J. stat. 1990, 18: 141~148
- [99] Shi X Q. Some asymptotic results for Jackknifing the sample quantile. Ann Stat. 1991, 19: 496~503
- [100] Singh D, Jindal K K, Garg J N. On modified systematic sampling. Biometrika, 1968, 55: 541~546
- [101] Singh R. Approximately optimal stratification on the auxiliary variable, Jour Amer Statist Assoc, 1971, 66: 829~833
- [102] Statistical Office of the United Nations. Elements of sample survey theory. 1972,
- [103] Stein C. A two-sample test for a linear hypothesis whose power is independent of the variance. Ann Math Stat, 1945, 16: 243~258
- [104] Stephan F F. The expected value and variance of the reciprocal and other negative powers of a positive Bernoulli variate. Ann Math Stat, 1945, 16: 50~61
- [105] Stephan F. McCarthy P J Sampling Opinions. New York, John Wiley and Sons, 1958, P. 243
- [106] Sukhatme P V. Sampling Theory of Surveys, With Application. Iowa State College Press, Ames Iowa. 1954
- [107] Tin M. Comparison of some ratio estimators. Jour. Amer. Stat. Assoc. 1965, 60: 294~307
- [108] Tschuprow A. A. On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. Metron, 1923, 2: 461~493, 646~683
- [109] Tukey J W. Bias and confidence in not-quite large Samples (Abstract). Ann Math. Statist. 1953, 29: 614
- [110] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias. Jour Amer Stat Assoc.
- [111] Wolter K M. Introduction to Variance 1965, 60: 63~69 Estimation. Springer Series in Statistics, 1985
- [112] Wu C F J. On the asymptotic properties of the Jackknife histogram, Ann Stat. 1990, 18: 1438~1452
- [113] Yates F. Systematic sampling. Phil. Trans. Roy. Soc. London, 1948, A241: 345~377
- [114] Yates F. Sampling Methods for Censuses and Surveys. Charles Griffin and Co., London, third edition, 1960
- [115] Yates F, Grundy P M. Selection without replacement from within strata with probability proportional to size. Jour Roy Stat Soc. 1953, B15: 232~261